Timing of *de novo* mutations

relevance to health and disease

Rocío Acuña Hidalgo

Timing of *de novo* mutations - relevance to health and disease

Rocío Acuña Hidalgo

Printing of this thesis was financed by the Radboud University Medical Center, Nijmegen, The Netherlands.

ISBN 978-94-6299-603-8

Printing: Ridderprint BV

Layout and cover design by Rocío Acuña Hidalgo

Cover image by Alex Lehner (https://goo.gl/eZrEry)

Timing of *de novo* mutations - relevance to health and disease

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 8 juni 2017
om 11.30 uur precies

door

Rocío Acuña Hidalgo
geboren op 14 februari 1985
te Buenos Aires, Argentinië

Promotor:
Prof. dr. ir. J.A. Veltman

Copromotoren:
Dr. A. Hoischen
Dr. C.F.H.A Gilissen

Manuscriptcommissie:
Prof. dr. A. Cambi
Prof. dr. C. Noordam
Dr. G.W.E. Santen (Leids Universitair Medisch Centrum)

Timing of *de novo* mutations - relevance to health and disease

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Thursday, June 8, 2017
at 11.30 hours

by

Rocío Acuña Hidalgo
born on February 14, 1985
in Buenos Aires, Argentina

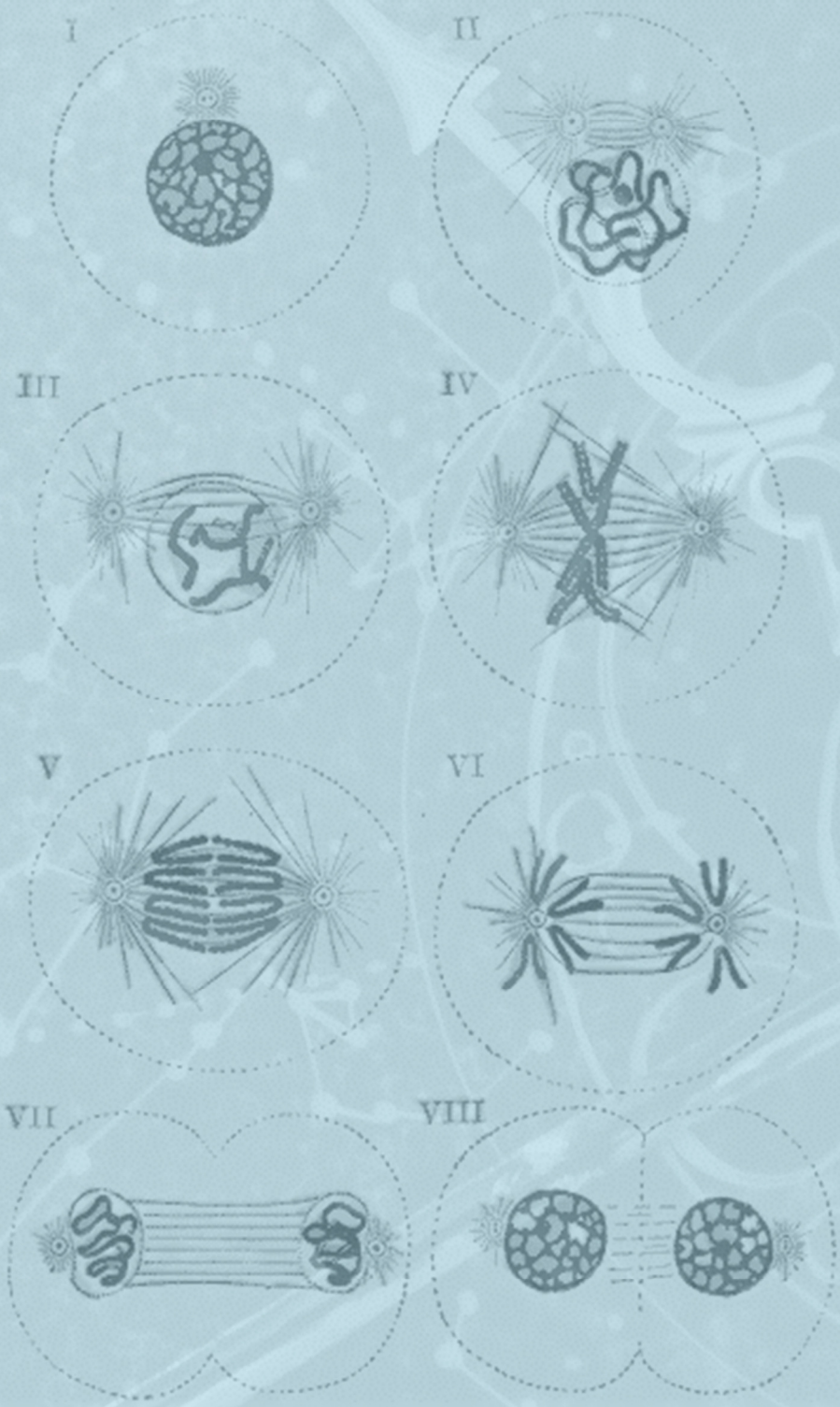
Supervisor:
Prof. dr. ir. J.A. Veltman

Co-supervisors:
Dr. A. Hoischen
Dr. C.F.H.A Gilissen

Doctoral Thesis Committee:
Prof. dr. A. Cambi
Prof. dr. C. Noordam
Dr. G.W.E. Santen (Leiden University Medical Center)

A mis padres

Table of contents



Mitotic division. I to III, prophase; IV, metaphase; V and VI, anaphase; VII and VIII, telophase. Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 1:
General introduction

Glossary

Germline mutation	Mutation that is present in the sperm or egg cell and is inherited to the offspring, who will then have the mutation in all the cells of its organism.
Somatic mutation	Mutation that occurs in a somatic cell in an organism, that is any cell of the organism other than the gametes. Somatic mutations are not inherited to the offspring.
Postzygotic mutation	Mutation that arises in a cell of an organism in the first few cell divisions after fertilization.
Mosaicism	Presence of more than one genetically distinct cell line in a single organism deriving from one zygote.
Chimerism	Presence of more than one genetically distinct cell lines in an organism deriving from different zygotes.
Primordial germ cell	Cell set aside early in embryogenesis that serves as a progenitor to the germ cells which eventually give rise to the sperm or egg cell.
Germ layer	Layer of cells formed during embryonic development by gastrulation. Vertebrate embryos have three germ layers: ectoderm, mesoderm and endoderm.
Germline mosaicism	Presence of mosaicism in the gonads, with some germ cells having a mutation while others do not.
Somatic mosaicism	Presence of mosaicism in an organism, excluding the gonads. Some cells in the organism have a mutation, while others do not.

DNA holds information that is essential for life

Every living organism on Earth carries in each of its cells the genetic information needed for its development, growth and maintenance.[1] In humans, this information is stored in the form of deoxyribonucleic acid (DNA), which provides cells with the instructions needed to produce all the building blocks necessary for life. Additionally, DNA allows for genetic information to be transmitted from a cell to its daughter cells, but also from an organism on to its progeny.

DNA is composed by long chains of nucleotides which consist of a sugar molecule and one of four bases: adenine (A), cytosine (C), guanine (G) or thymidine (T). DNA is most often found in the form of a double helix, which results from the binding of two nucleotide chains to one another.[2] Nucleotides are complementary and pair in only two ways: A with T and C with G.[3] Base-pair complementarity implies that the sequence of nucleotides in a single strand of DNA is sufficient to copy and transmit all genetic information. DNA can be replicated by separating a double stranded molecule of DNA and using each strand as a template to which complementary base pairs are added by a DNA polymerase to form a second strand of DNA.[2,4] Based on the sequence of the nucleotides in each strand, DNA encodes the genetic information that serves as the template to make proteins.[5]

A genome is the complete set of genetic information of an organism, including both nuclear or genomic DNA (gDNA) and mitochondrial DNA. In humans, each cell contains approximately 6 billion base pairs of gDNA ($6x10^9$ bp), which would equal roughly 3 meters long if extended.[6] However, gDNA is tightly packaged by wrapping itself around proteins to form chromatin (Figure 1), which is then stacked and compacted to form chromosomes.[7] Humans have 46 chromosomes forming two sets of 23 chromosomes, of which one pair corresponds to sex chromosomes that differ between the sexes: XX in women and XY in men. In contrast, the remaining 22 pairs, also called the autosomes, are present regardless of the sex. Almost all cells in the human body are diploid: they contain two sets of 23 chromosomes in their nucleus of which one set is inherited

from the mother and one from the father. Germ cells, such as the sperm or egg cells, are an exception as they have a single set of chromosomes and are therefore haploid.

The central dogma of molecular biology: DNA to RNA to protein

Chromosomes contain genes, which are segments of gDNA with genetic instructions to make a functional RNA molecule or a protein. As of 2016, the human genome is estimated to contain approximately 20,000 protein-coding genes, accounting only for 1% of the human genome.[6] Base-pair complementarity is also fundamental for the expression of information contained in DNA. Gene expression begins with the separation of double stranded DNA in two strands, one of which is then used as a template for the addition of nucleotides by an RNA polymerase to form messenger RNA (mRNA).[8] This process of forming an mRNA molecule by copying a gene is termed RNA transcription and is similar to DNA replication, except that thymine is replaced with uracil, which also pairs with adenine. The mRNA transcript is then processed by the splicing machinery in the cell nucleus to remove the introns, which are regions of a gene that are transcribed from DNA into mRNA but are not included in the mature mRNA. The exons, which are the regions of a gene remaining after splicing, are then joined to form mature mRNA.[1]

Mature mRNA is exported from the nucleus to the cytoplasm or to the membrane of the rough endoplasmic reticulum, where the sequence of nucleotides is read and translated into protein by the ribosome. The genetic code is the basis for translation of the mRNA transcript into protein; each codon, or combination of three successive nucleotides, encodes for one of 20 different amino acids or for one of three signals to stop translation of the mRNA into protein (stop codon).[1] RNA consists of four different nucleotides (A, C, G and U) which give rise to 64 different codons. Transcripts are read by dividing the sequence of DNA into triplets of nucleotides which match one of the 64 codons. Any transcript can be read in three different frames, beginning from each of the three nucleotides composing a triplet. Usually, only one reading frame leads to the correct production of a protein (the open reading frame). Since there are only 20 different amino acids, there is redundancy in the genetic code and different codons can code for the same amino acid (synonymous codons). The transcript is therefore read one codon at a time, deciphered and the corresponding amino acid is incorporated into the protein before moving on to the next codon. Different amino acids have distinct chemical properties and the interaction between the side chains of a sequence of amino acids within a protein determines how the translated protein folds to form protein secondary structures, such as alpha helixes and beta sheets. Proteins then fold into a three dimensional structure which is necessary for its correct function.[9] Proteins can be modified by

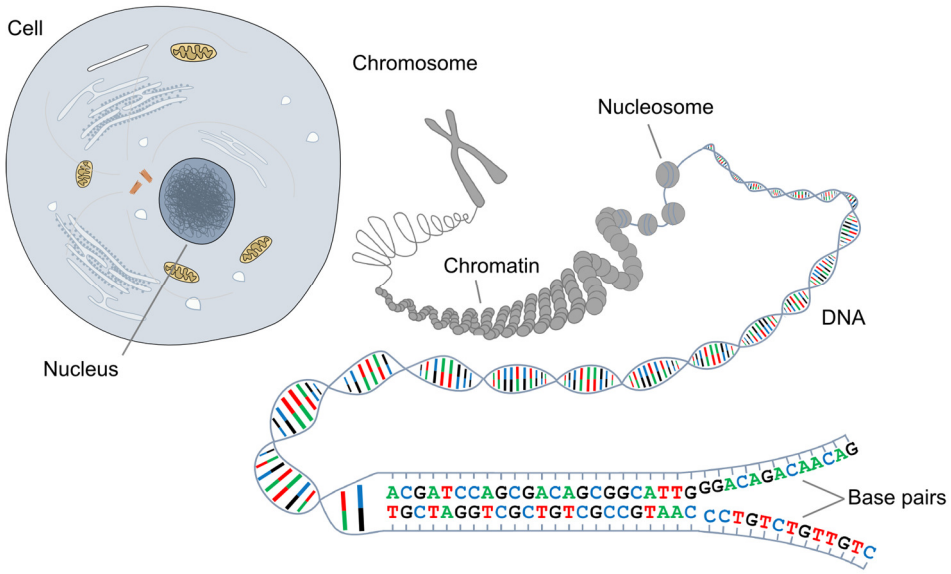


Figure 1. Depiction of a cell and multiple DNA structures. DNA is formed by four base pairs (A, C, G and T) which are complementary to each other and form chains of nucleotides. DNA is wrapped around proteins to form nucleosomes. Nucleosomes are packaged together to form chromatin, which is in turn compacted with scaffolding proteins into solenoid loops. During cell division, chromosomes become more condensed to make chromosome segregation possible.

addition of chemical molecules such as a phosphate group or a saccharide during or after protein synthesis in the cell.

DNA that does not code for proteins (non-coding DNA) may still be functional; it may contain regulatory elements such as enhancers, repressors, silencers, promoters and transposons that are not transcribed into RNA.[10] Additionally, non-coding DNA may also include genes that are transcribed into RNA molecules which are functional by themselves, such as ribosomal RNA or microRNA.[11]

Genetic variation is at the origin of evolution, diversity and genetic diseases

Genetic information is encoded in DNA by the sequence of nucleotides; the genome of an individual organism dictates not only the general characteristics of its species but also the personal variation pertaining to that specific individual. The human reference genome is a digitally built DNA sequence representing the genome of the human species (*Homo sapiens*). The human reference genome differs substantially in size and in sequence from the reference genome of other organisms, such as the thale cress (*Arabidopsis*

thaliana)[12] or the mouse (*Mus musculus*).[13] However, there are also many differences in the sequence of nucleotides in the genome of any human being when compared to the human reference genome.

The sequence of nucleotide bases in the genome can mutate, due to errors in the normal mechanism by which genomes are copied or repaired when damaged. Alterations in the nucleotide sequence resulting from these mutations can become permanently fixed in a genome sequence, giving rise to variation within a genome. Mutations leading to genetic variation can occur at the level of a single nucleotide, in which case they are called single nucleotide variants (SNVs), or they can affect multiple bases. One or several nucleotides can be inserted or deleted in a fragment of DNA, which is known as an indel if the size of the insertion or deletion is less than 1000 bp or 1 kb. Variation affecting DNA fragments larger than 1 kb is known as structural variation (SV) and includes insertions and deletions but also duplications, inversions and translocations.

Differences in the sequence of A, C, G and T lead to variation in the information encoded genetically; if a SNV occurs in the coding region of a gene, it may affect translation of the protein which may have consequences on the functionality of the protein. For instance, a SNV in a codon can lead to the incorporation of a different amino acid during translation (missense mutation). A SNV replacing a T for a G resulting in a change in the codon from UUG to UGG would lead to the substitution of the encoded residue from leucine to tryptophan. Because of the different chemical properties of leucine and tryptophan, this missense mutation can have important consequences on the structure of the protein and its function. On the other hand, because of the redundancy of the genetic code, a SNV may not lead to any changes in the sequence of amino acids (synonymous mutation). A substitution of a G to an A leading to a codon change from UUG to UUA would not have an effect on the translation of the protein, since UUG and UUA both code for leucine. A SNV can also lead to the substitution of a codon by a stop codon (nonsense mutation). For example, a change of a T to an A leading to a codon substitution of UUG for UAG would be read by the ribosome as an indication to stop translating the transcript into protein, which leads to a truncation in the amino acid sequence. Finally, because the translation occurs codon by codon, the insertion or deletion of a base pair may change the entire reading frame of the transcript (frameshift mutation). This can ultimately also lead an early truncation of the protein or the translation of a protein that is not functional.

Two unrelated humans share approximately 99.9% of their genome sequence. The remaining 0.1% of the human genome that is variable between individuals includes both common and rare variants. A typical human genome varies at 4.1 to 5.0 million positions compared to the human reference genome.[14] The vast majority of genetic variation observed in a typical human genome is common and shared by more than 0.5% of the population as a result of having been recombined, selected and passed on for many generations.[14] On the other

hand, a typical human genome contains 40,000 to 200,000 rare variants that are observed in less than 0.5% of the population.[14] A recent analysis of the coding sequence of 60,000 individuals showed that 99% of all variants identified in humans have a population frequency of less than 1%.[15] This study identified 7.4 million common and rare variants in the coding sequence of the human genome, which corresponds to one variant every 8 bp of coding sequence.[15] A gene may have different mutations throughout its sequence and each alternative form of a gene or a genetic locus is known as an allele. Genes occur in pairs in the human genome because we have two sets of chromosomes, implying that humans have two alleles for each gene. If both alleles of a particular gene are identical in an organism, then the organism is said to be homozygous for this allele. Alternatively, if the two alleles are different, then the organism is heterozygous for said allele.

Genetic variation in functional regions of the genome may cause changes in gene expression or in the sequence and structure of these encoded RNA or protein molecules, which in turn leads to differences in their function. Changes in the function of a specific RNA or protein molecule can result in variation in a trait, a specific feature which may range from a molecular aspect, such as intolerance to certain types of food because of an enzymatic defect, or to a physical characteristic, like hair color. Seemingly small genetic changes may give rise to significant physical or physiological differences. For instance, a single nucleotide change in a non-coding region of the human genome associated with blonde hair color in European populations was recently shown to disrupt the expression in hair follicles of *KITLG*, a gene involved in migration, survival and proliferation of melanocytes.[16,17]

Genetic variation, together with the environment, determine the sum of an organism's traits, or its phenotype. Phenotypes in humans are incredibly diverse and can vary both in physiologic and pathogenic ways. Genetic and phenotypic variation can affect the fitness of an organism, which is the likelihood of an organism to survive and reproduce in a given environment. The environment is often changing and the fitness of a given allele is likely to be unstable and fluctuate through time in response to the environment.[18] Natural selection is the process by which a certain trait results in differences in fitness in individuals with this trait, leading to differential survival and reproduction of the allele responsible for the trait.[19]

Mutations with a survival or reproductive advantage spread rapidly through a population. Recent studies in human populations living in extreme environmental conditions or with significant dietary changes have shown evidence of natural selection for specific alleles and traits.[20] For instance, genetic variation in a transcription factor involved in response to hypoxia shows a strong signal for natural selection in ethnic Tibetans living at high altitudes, suggesting adaptation to the low oxygen found in this environment.[21] Similarly, strong

Inheritance pattern	Mechanism	Examples
Autosomal recessive	The phenotype is caused by the presence of a disease-causing allele in both copies of a gene, either with the same allele on both copies or a different allele in each copy.	Cystic fibrosis (*CFTR*)[35]
Autosomal dominant	The phenotype is caused by the presence of a disease-causing allele in one copy of a gene.	Marfan syndrome (*FBN1*)[36]
X-linked	The phenotype is caused by the presence of a disease-causing allele on the X chromosome.[a]	Duchenne's muscular dystrophy (*DMD*)[37]
Germline *de novo* mutations	The phenotype is caused by a germline mutation in the affected individual which was not inherited from the parents and thus likely occurred during gametogenesis.	Bohring-Opitz syndrome (*ASXL1*)[26]
Postzygotic *de novo* mutations	The phenotype is caused by a mutation in the affected individual which occurs in the first few cell divisions after fertilization.	Proteus syndrome (*AKT1*)[38]
Digenic or oligogenic	The phenotype is caused by damaging variants in two or more genes.	Bardet-Biedl syndrome (*BBS2, BBS6)*[39]
Imprinting disorders	The expression of the phenotype depends on the parent of origin of the pathogenic allele due to parent-specific gene expression.	Beckwith-Wiedemann syndrome (*BWS*)[40]
Complex disorders	The phenotype is caused by the combination of a multigenic component and environmental factors.	Alzheimer's disease (*APOE*)[41]
Repeat expansions	Expansion of repeats of trinucleotides due to genetic instability.	Huntington's disease (>40 CAGs in *HTT*)[42]
Mitochondrial genetic disease	Mutations in genes within the mitochondrial genome.	Leigh disease (*MT-ATP6*)[43]

Table 1. Mendelian and non-Mendelian modes of inheritance. [a] X-linked disorders can also be recessive or dominant, depending on whether the presence of a wild-type allele in females with two X chromosomes can rescue the effect of the disease-causing allele.

selection for alleles resulting in the persistence of lactase activity in adulthood have been identified in European population, where additional nutrition from dairy consumption may have given a selective advantage.[22] However, high frequency of a mutation within a population does not always entail that it is under positive selection; most genetic variation is neutral and does not result in phenotypic differences. In these cases, the frequency of a neutral allele can change simply because of chance, without undergoing natural selection, an effect known as genetic drift.[19]

Genetic variation can result in deleterious traits and phenotypes which may be disadvantageous in the context of a specific environment or which constitutively impair the normal functioning of an organism. From the perspective of evolution, individuals with deleterious traits have decreased survival or reproductive success, which progressively leads to the elimination of damaging mutations from the population. This phenomenon is known as purifying selection and in the most extreme cases, an affected individual may die without producing progeny, entirely removing the damaging mutation from the population. From a medical perspective, an individual with mutations in DNA disrupting his or her development, physiology or anatomy is considered to have a genetic disease. Some genetic diseases are thought to result from the interaction of variation in several genes together with lifestyle and environmental factors. These disorders, for which no clear inheritance pattern has been identified, are therefore known as multifactorial or complex disorders (Table 1). On the other hand, some genetic disorders have been established to be caused by mutations in one gene or genetic locus. These diseases often affect many members within a family and are known as monogenic or Mendelian disorders.

Monogenic disorders are a prominent cause of human disease

Monogenic disorders are individually rare; the most common monogenic disorder in Northern European population is cystic fibrosis, a disease caused by mutations in *CFTR* and found in 1 in 2000 births.[23] Collectively, however, monogenic disorders are estimated to affect approximately 0.4% of live births,[24] representing a substantial part of human disease and placing an important burden on healthcare systems worldwide.

Like all genetic information, damaging mutations leading to genetic disorders can be inherited from parents to offspring and monogenic diseases can be classified as recessive or dominant disorders according to their mode of inheritance. Diploid cells have two copies of each gene and, for most genes, loss of function in one of the copies due to a damaging mutation can be compensated by the presence of a normal allele on the other chromosome. Thus, the function of most genes is only disrupted if both copies harbor mutations, either because the same mutation is found homozygously or if different mutations are present on each copy (compound heterozygous). Monogenic disorders requiring the presence of homozygous or compound heterozygous mutations are recessive disorders (Table 1 and Figure 2A).

Dominant diseases arise in the presence of only one mutated allele and their molecular mechanism of disease is variable. In some cases, the loss or damage to one copy of the gene cannot be compensated by the remaining one due to a dosage problem (haploinsufficiency). In other instances, the mutated

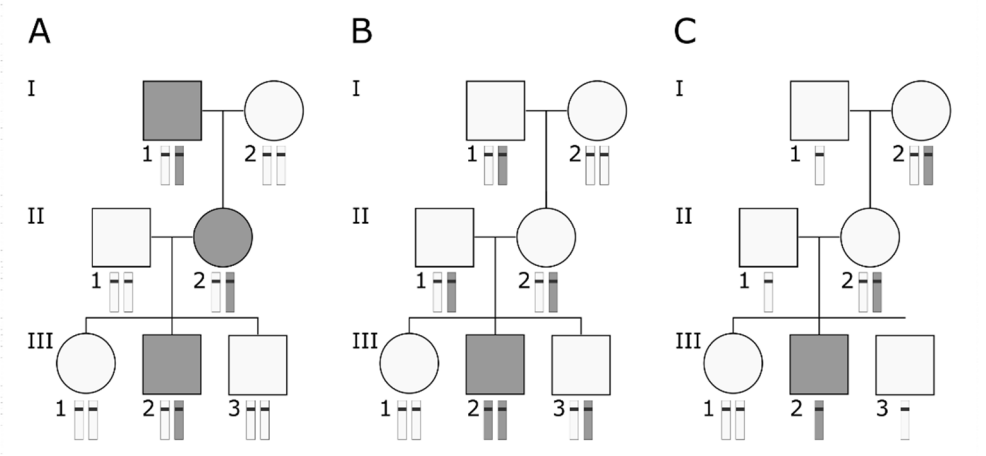


Figure 2. A. Family tree showing the transmission of an autosomal dominant pathogenic allele. Individual I-1 carries a pathogenic allele and is affected, transmitting the allele and the disease to his daughter II-2, who in turn transmits it to her son III-2. B. Family tree showing the transmission of an autosomal recessive pathogenic allele. Individual I-1 carries the pathogenic allele but is not affected by the disease. He transmits the pathogenic allele to his daughter II-2 who has children with another unaffected carrier (II-1). Among her three children, one is affected due to inheritance of both pathogenic alleles (III-2) and one is an unaffected carrier (III-3). C. Family tree showing the transmission of an X-linked pathogenic allele. Individual I-2 carries the allele but does not display the phenotype. She transmits the pathogenic allele to her daughter II-2 who is unaffected, but who in turn transmits the allele to her son III-2, who is a carrier of the allele and displays the phenotype.

allele may interfere with the function of the normal allele, causing dysfunction of both copies (dominant negative mutation). Finally, other mutations may change the function of the allele (gain-of-function mutation), which cannot be compensated by the normal allele. Dominant disorders are usually transmitted by an affected parent in which the mutation is present heterozygously (Table 1 and Figure 2B). However, dominant disorders can also be caused by *de novo* mutations, which are mutations arising in the germline for the first time. *De novo* mutations have been recently recognized as an important cause of dominant disorders, particularly in sporadic cases.

In broad terms, genetic disorders caused by mutations in genes located on the autosomes occur independently of the sex of the individual. However, because some genes are found on the sex chromosomes, the inheritance of monogenic diseases can also be linked to the sex of an individual. The vast majority of disease genes in the sex chromosomes are found on chromosome X. Genetic disorders caused by mutations in these genes are said to have X-linked inheritance and rarely affect women, as the presence of a second copy of chromosome X in women often rescues a pathogenic mutation. On the other hand, as a result of having a single copy of the X chromosome, men are more

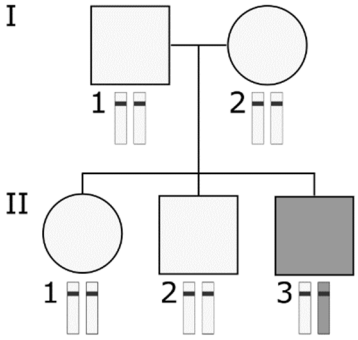


Figure 3. Family tree showing germline *de novo* occurrence of a mutation. Both parents have wild-type alleles and have two children with wild-type alleles and one with a constitutive *de novo* mutation.

susceptible to X-linked disorders and are usually more severely affected (Table 1 and Figure 2C). Few disease genes, implicated for the most part in infertility, have been identified in chromosome Y and consequently only affect men.

Many disorders with a clear genetic component do not follow Mendelian inheritance patterns. A prominent example of this are some genetic disorders with sporadic appearance, where both the disease-causing allele and the phenotype are present in the offspring but absent in the parents. As a result, the main mechanism for these disorders is the *de novo* occurrence of a pathogenic mutation in the offspring, leading to the disease phenotype (Figure 3). Some sporadic diseases caused by *de novo* mutations may not affect fertility and affected individuals can transmit the mutation and the disorder to their offspring as an autosomal dominant disease.[25] Other phenotypes, however, occur exclusively due to *de novo* mutation because the phenotype entails decreased fitness and fertility, which precludes the transmission of the pathogenic allele.[26–28]

The origin and role of *de novo* mutations in human biology and disease are reviewed in more detail in **Chapter 2**. A few additional examples of disorders with non-Mendelian segregation, such as somatic mutations, complex diseases and mitochondrial genetic disorders are summarized in Table 1.

The next generation sequencing (NGS) revolution

Many genetic techniques have been developed to detect genetic mutations, including karyotyping, microarray technology, Sanger sequencing and more recently, next generation sequencing (NGS). An important distinction between these methods is their level of resolution, which is the minimal size of the alteration that can be detected using this technique. Genetic disease in

humans can be caused by mutations ranging from an alteration in a single nucleotide to the involvement of an entire chromosome. Consequently, an appropriate genetic technique with the right level of resolution must be selected to be able to detect each of these types of variation (see Table 2). For instance, karyotyping can only visualize aneuploidies and large chromosomal aberrations, while Sanger sequencing can detect small indels and SNVs at known loci.

In contrast with these methods, NGS has the capacity to detect all kinds of genomic variation. NGS is based on the massive sequencing of millions of DNA fragments in parallel, offering an unprecedented output of genetic information at the base pair level in a single experiment. Many different high throughput sequencing methods exist based on distinct principles to generate millions of sequence reads in parallel. Currently, the sequencing methods most commonly used generate short sequencing reads of 35 to 700 base pairs. These are based mainly on two principles: sequencing by synthesis (with cyclic reversible termination as used by Illumina or single nucleotide addition such as in semiconductor sequencing in the Ion Torrent system and pyrosequencing in the 454 Pyrosequencing, both by Life technologies) and sequencing by ligation (used by Life technologies in the SOLiD systems and by Complete Genomics in DNA nanoball sequencing).[29,30] Single molecule real time sequencing (used by Pacific Biosciences) and nanopore sequencing (used by Oxford Nanopore Technologies) are two promising sequencing methods producing long reads of up to 200 kb[29]. In addition to the sequencing approach and read length, these NGS methods have different accuracy, throughput, costs and, therefore, applications. The increased accuracy of NGS-based approaches relies on high sequencing coverage, which is the number of individual reads covering a genomic region.

Despite the fact that the advent of NGS entails an exponential increase in the sequencing throughput of these platforms, sequencing of entire eukaryotic genomes is not yet routinely feasible due to costs and time. Whole genome sequencing (WGS) of one human genome at 30x coverage represents 180 Gb of sequencing, which until relatively recently required several sequencing runs and thousands of dollars in costs.[31] As a result, a lot of effort has been placed in developing methods to target regions of interest within a genome for sequencing. Specific genomic regions can be enriched by multiplexed PCR, capture by molecular inversion probes (MIPs) and hybrid selection either in solution or in an array.[31] For instance, whole exome sequencing (WES) is a method to enrich for the exome, that is all the exons of the genome which represent approximately 1% of the genome. In order to enrich for this fraction, genomic DNA is randomly sheared into smaller fragments which are subsequently hybridized to a biotin-labelled probe that is specific to the region targeted. Streptavidin is then used to pull out the biotinylated probes and extract the bound DNA fragments. These enriched regions are then processed into a sequencing library that is read by a NGS platform of choice. Thus, for a single

	Karyotype	Microarray	Sanger sequencing	NGS
Basis of the technique	Staining of chromosomes and observation under a light microscope	DNA hybridization onto a microarray with large insert clones or oligonucleotide probes	PCR amplification of a target sequence incorporating fluorescent dideoxynucleotides that terminate DNA synthesis	Massive parallel sequencing of millions of DNA fragments by amplification, ligation or other sequencing principles
Genome-wide or targeted technique	Genome-wide	Genome-wide	Targeted	Genome-wide or targeted
Resolution of the technique	Aneuploidies and chromosomal aberrations (>5 Mb)	1 to 100 kb, depending on the array	Single nucleotide	Single nucleotide
Type of variation that can be detected	Gains, losses, rearrangements	Gains, losses	SNVs, indels	Gains, losses, indels, SNVs
Example of genetic alteration that can be detected	Trisomy 21 in Down's syndrome[44]	8q12 microdeletion in CHARGE syndrome[45]	Inherited *FGFR3* point mutation in achondroplasia[46]	*De novo* SNVs, indels, copy number variations and chromosomal rearrangements in intellectual disability[47]
Comment	Limited resolution	Fails to detect balanced rearrangements	Low throughput, time-consuming and expensive	High throughput, need for bioinformatics infrastructure, challenges in variant interpretation.

Table 2. Comparison of techniques to detect genetic alterations.

individual, exome-wide variation can easily be detected by WES. On the other hand, MIPs represent a capture method that allows multiplexing and scaling so that one can target a set of genomic regions of interest and easily sequence them in hundreds of samples in a single NGS platform run. In addition to enrichment methods, the protocol for library preparation can be modified in order to examine other aspects of cell biology, such as the transcriptome, the epigenome or even study single cells.

NGS allows for genome-wide detection of genetic variation at nucleotide level. As a result, NGS has fueled a revolution in human genetics, particularly

within fields of research that had been previously limited by the available technology. For instance, traditional disease gene identification methods are based on Sanger sequencing to identify disease-causing mutations in candidate genes. This approach relies on the identification of candidate genes by positional mapping of loci in large pedigrees with multiple members affected or based on previous knowledge of the biological or genetic aspects of the disease.[32] On the other hand, WES and WGS provide the possibility to perform an untargeted and therefore unbiased exome- or genome-wide analysis in a single affected individual.

In the first 5 years of widespread availability of WES, more than 500 genes associated with monogenic disorders have been identified by WES and WGS, of which an important proportion involve *de novo* mutations arising for the first time.[24,32,33] Trio-based sequencing represents one of the stepping stones for discovery of *de novo* mutations and identification of candidate disease-genes. Trio-based sequencing consists of performing WES or WGS in an individual and both his or her parents. This allows the identification of mutations that were not inherited but occurred *de novo* and are thus are present in the offspring while absent in the parents.[32] This approach relies heavily on the quality of the sequencing method and the accuracy of the software used to call variants.[32] A large fraction of mutations identified as *de novo* by trio-based sequencing are in fact sequencing artifacts or inherited mutations missed in one or both parental samples.[34]

The generation and role of *de novo* mutations in human health and disease is a fundamental topic in my thesis and, as such, **Chapter 2** serves as an in-depth introduction to this topic. Additionally, in this chapter, I provide additional insights into NGS and somatic mutations.

References

1. Alberts, B. Johnson, A. Lewis, J. Raff, M. Roberts, K. Walter, P. *Molecular Biology of the Cell, 5th Edition*. *Garland Science* (2008).
2. Watson, J. & Crick, F. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–8 (1953).
3. Watson, J. & Crick, F. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171,** 964–7 (1953).
4. Kornberg, A. Biologic Synthesis of Deoxyribonucleic Acid. *Science* **131,** 1503–1508 (1960).
5. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–3 (1970).
6. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
7. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98,** 285–294 (1999).
8. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190,** 576–581 (1961).
9. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223–30 (1973).
10. Riethoven, J.-J. M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol* **674,** 33–42 (2010).
11. Mattick, J. S. Non-coding RNA. *Hum Mol Genet* **15,** R17–R29 (2006).
12. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408,** 796–815 (2000).
13. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
14. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
15. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
16. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* **39,** 1443–1452 (2007).
17. Guenther, C. a, Tasic, B., Luo, L., Bedell, M. a & Kingsley, D. M. A molecular basis for classic blond hair color in Europeans. *Nat Genet* **46,** 748–52 (2014).
18. Orr, H. A. Fitness and its role in evolutionary genetics. *Nat Rev Genet* **10,** 531–539 (2009).
19. Fu, W. & Akey, J. M. Selection and Adaptation in the Human Genome. *Annu Rev Genomics Hum Genet* **14,** 467–489 (2013).
20. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349,** 1343–1347 (2015).
21. Yi, X. *et al.* Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329,** 75–78 (2010).
22. Bersaglieri, T. *et al.* Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *Am J Hum Genet* **74,** 1111–1120 (2004).
23. Kerem, B. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245,** 1073–1080 (1989).
24. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97,** 199–215 (2015).
25. Rousseau, F. *et al.* Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. *Nature* **371,** 252–4 (1994).
26. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* **43,** 729–731 (2011).
27. Rivière, J.-B. *et al.* De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. *Nat Genet* **44,** 440–4, S1-2 (2012).
28. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42,** 790–793 (2010).
29. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).

30. Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11,** 31–46 (2010).
31. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7,** 111–8 (2010).
32. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* **20,** 490–497 (2012).
33. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14,** 681–91 (2013).
34. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12,** 745–755 (2011).
35. Rommens, J. *et al.* Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science (80-)* **245,** 1059–1065 (1989).
36. Dietz, H. C. *et al.* Marfan syndrome caused by a recurrent de novo missense mutation in the fibrillin gene. *Nature* **352,** 337–339 (1991).
37. Koenig, M. *et al.* Complete cloning of the duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50,** 509–517 (1987).
38. Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* **365,** 611–9 (2011).
39. Katsanis, N. *et al.* Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* **293,** 2256–9 (2001).
40. Hatada, I. *et al.* An imprinted gene p57KIP2 is mutated in Beckwith–Wiedemann syndrome. *Nat Genet* **14,** 171–173 (1996).
41. Bertram, L. & Tanzi, R. E. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nat Rev Neurosci* **9,** 768–778 (2008).
42. Pearson, C. E. Slipping while sleeping? Trinucleotide repeat expansions in germ cells. *Trends Mol Med* **9,** 490–5 (2003).
43. Wallace, D. C. Mitochondrial diseases in man and mouse. *Science (80-)* **283,** 1482–1488 (1999).
44. Lejeune, J., Gautier, M. & Turpin, R. Etude des chromosomes somatiques de neuf enfants mongoliens. [Study of somatic chromosomes from 9 mongoloid children]. *Comptes rendus Hebd des se• ances l'Acade• mie des Sci* **248,** 1721–2 (1959).
45. Vissers, L. E. L. M. *et al.* Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat Genet* **36,** 955–957 (2004).
46. Shiang, R. *et al.* Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. *Cell* **78,** 335–342 (1994).
47. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).

Aims and scope of this thesis

The introduction and expansion of the use of NGS to study the genetics of human disease in the last years swiftly removed the restrictions imposed by the limitations of previous technologies. These developments have led to significant progress in the study of *de novo* mutations as a cause for sporadic human disease. This includes both *de novo* germline mutations, which often play a crucial role in the developmental disorders, and new somatic mutations, which often drive different forms of cancer. Remarkably, some germline mutations causative for developmental disorders have been shown to arise in the same genes as somatic mutations driving cancer. While mutations disrupting the same gene can cause both types of disease, the downstream effects of the mutation and the expression of the phenotype are likely modulated by the type and timing of the mutation. Therefore, the aim of this thesis is to study the timing of *de novo* mutations in human physiology and human disease, particularly for mutations which may result either in developmental disorders or in cancer.

In **Chapter 2**, I describe the causes and mechanisms of *de novo* mutations and expand on their role in human biology and disease.

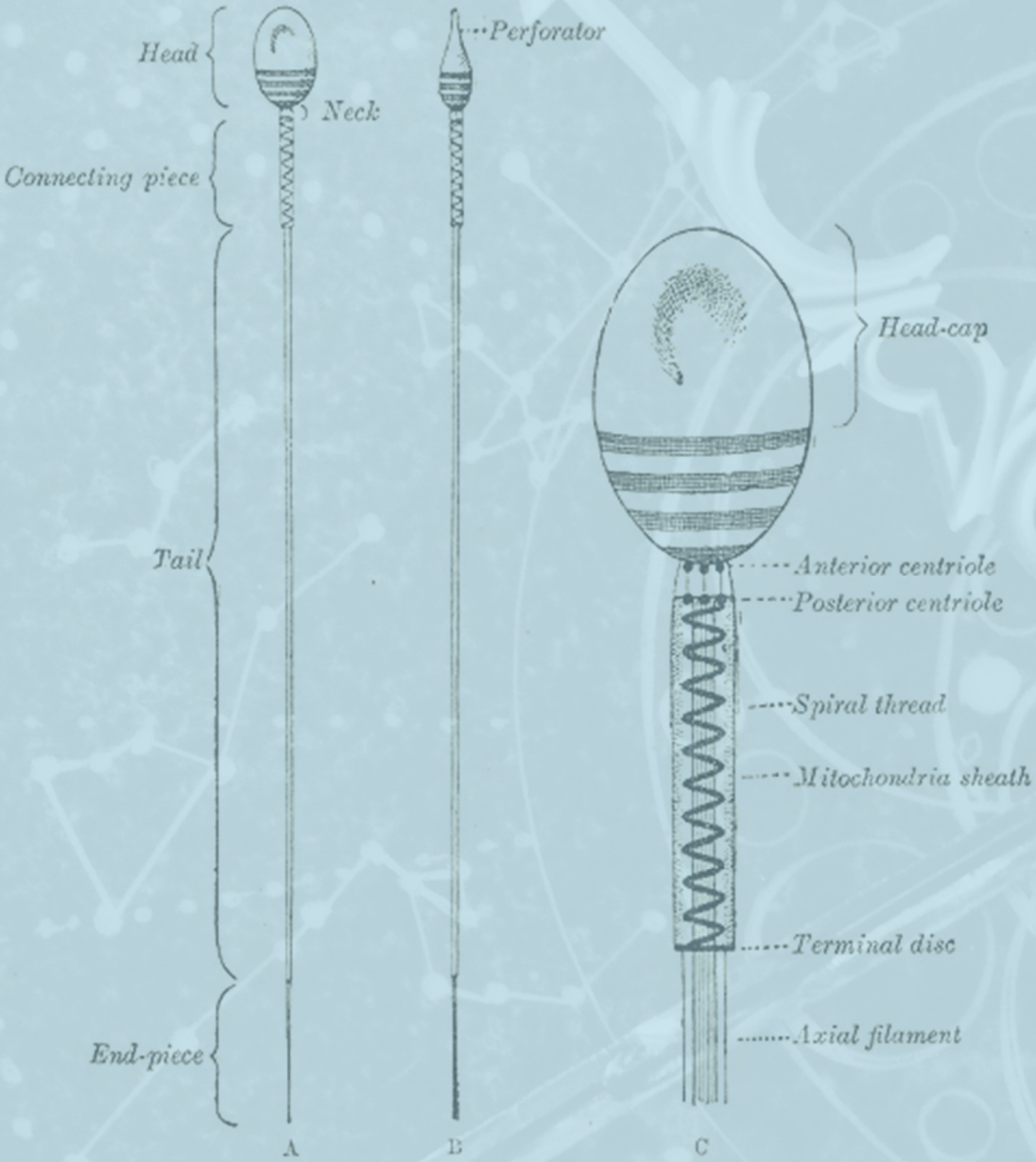
Chapter 3 presents an instance in which the application of NGS led to the identification of the disease-causing gene for a sporadic phenotype caused by *de novo* mutations. In this chapter, I describe how we applied exome sequencing to detect *de novo* mutations in *THRA* as the cause of a novel thyroid hormone resistance syndrome.

Chapter 4 shows that the use of NGS can expand beyond novel disease-gene identification, by offering insight into the timing *de novo* mutations. By analyzing the variant allele fraction of *de novo* mutations sequenced using NGS methods, we detect *de novo* mutations which actually occurred as postzygotic events.

In **Chapter 5**, we investigate the timing of novel mutations throughout adult life by using human hematopoiesis as an *in vivo* model.

Chapter 6 explores the differences in timing and functional consequences of overlapping germline and somatic *SETBP1* mutations involved in a rare developmental disorder and in leukemia, respectively.

Chapter 7 provides the general discussion of this thesis and implications for future research on the topic.



Human sperm cell. A. Surface view. B. Profile view.
C. Magnification of the head, neck and connecting piece.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 2:
New insights into the generation and role of *de novo* mutations in health and disease

Adapted from:

Acuna-Hidalgo R., Veltman J.A. & Hoischen A. New insights into the generation and role of *de novo* mutations in health and disease. Genome Biol **17**, 241 (2016).

Abstract

Aside from inheriting half of the genome of each of our parents, we are born with a small number of novel mutations that occurred during gametogenesis and postzygotically. Recent genome and exome sequencing studies of parent-offspring trios have provided the first insights into the number and distribution of these *de novo* mutations in health and disease, pointing to risk factors which increase their number in the offspring. *De novo* mutations have been shown to be a major cause of severe early-onset genetic disorders such as intellectual disability, autism spectrum disorder and other developmental diseases. In fact, the occurrence of novel mutations in each generation explains why these reproductively lethal disorders continue to occur in our population. Recent studies have also shown that *de novo* mutations are predominantly of paternal origin and that their number increases with advanced paternal age. In this chapter, we provide an overview of recent literature on *de novo* mutations, covering their detection, biological characterization and medical impact.

Introduction

Upon fertilization, a human zygote inherits half of its genome from the mother via the oocyte and the other half from the father through the sperm. In addition to the genetic information passed on from generation to generation, each of us is born with a small number of novel genetic changes, *de novo* mutations, that occurred either during the formation of the gametes or postzygotically.[1,2] Additionally, novel mutations continue arising throughout post-natal and adult life both in somatic and germ cells. Only mutations present in the germ cells can be transmitted to the next generation.[3]

There is a long-standing interest in the study of the frequency and characteristics of *de novo* mutations in humans, as these are crucial to the evolution of our species and play an important role in disease. A typical human genome varies at 4.1 to 5.0 million positions compared to the human reference genome.[4] The vast majority of genetic variation observed in a typical human genome is common and shared by more than 0.5% of the population as a result of having been recombined, selected and passed on for many generations.[4] On the other hand, a typical human genome contains 40,000 to 200,000 rare variants that are observed in less than 0.5% of the population.[4] All of this genetic variation must have occurred as a *de novo* germline mutation in an individual at least once in human evolution.[5] Historically, the germline mutation rate in humans has been calculated by analyzing the incidence of genetic disorders; in 1935, Haldane estimated the mutation rate per locus per generation based on the prevalence of hemophilia in the population.[6,7] More recently, in 2002, Kondrashov accurately calculated the *de novo* mutation rate in humans by examining the mutation rate at known disease-causing loci.[8] Nowadays, next generation sequencing (NGS) approaches in parent-offspring trios can be used to directly study the occurrence of all types of *de novo* mutations throughout the genome, from single nucleotide variants (SNVs) to small insertions-deletions (indels) and larger structural variations. Genome-wide NGS studies place the germline *de novo* mutation rate for SNVs in humans at 1.0 to 1.8 $\times 10^{-8}$ per nucleotide per generation,[1,9–13] with

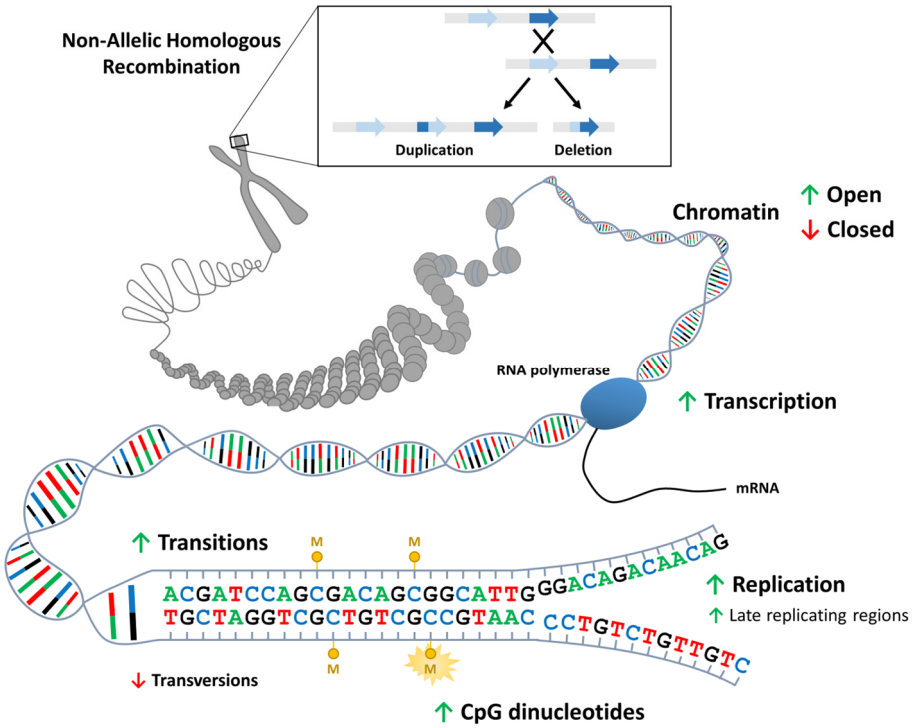


Figure 1. Mechanisms of *de novo* mutations. *De novo* mutations may arise because of static properties of the genome, such as the underlying sequence (deamination of methylated CpGs, transitions versus transversions) or due to erroneous pairing of nucleotides during DNA replication. However, *de novo* mutations may also occur in relation to cell-specific properties such as chromatin state, transcriptional status and gene expression levels. Mutational hotspots for genomic rearrangements are largely determined by the underlying genomic architecture. One such example is given for Non-Allelic Homologous Recombination. Arrows represent the influence of each feature on the *de novo* mutation rate. Green upwards arrows indicate elevated mutability; red downwards arrows indicate lower mutability.

substantial variation among families.[11,13,14] This number translates into 44 to 82 *de novo* single nucleotide mutations in the genome of the average individual, with 1 to 2 affecting the coding sequence.[9,10,12,13,15] These state-of-the art genomic approaches allow us to determine additional characteristics of *de novo* mutations, such as the parental origin and whether they occurred in the germline or postzygotically. We now know that the majority of germline *de novo* mutations have a paternal origin and that a higher paternal age at conception results in an increase in the number of *de novo* mutations in the offspring.[15–18] Furthermore, the study of large cohorts of parent-offspring trios provides insight into the distribution of mutations throughout the genome, the genomic context in which they arise and possible underlying mechanisms (see Figure 1 for an overview of different mechanisms resulting in *de novo* mutations).[11–13]

Mutations conferring a phenotypic advantage propagate rapidly through a population,[19–21] while neutral mutations may disseminate merely as a result of genetic drift.[22] However, damaging mutations resulting in deleterious traits before or during the reproductive phase undergo purifying selection and their spread through the population is averted.[23] This entails that *de novo* mutations are genetically distinct from inherited variants, as they represent the result of the mutagenic processes taking place between one generation and the next, prior to undergoing selection (see Table 1). Loss or acquisition of traits at the population level drives evolution of a species whereas, at the level of an individual, loss or acquisition of traits may result in disease.

Germline *de novo* genetic alterations have been implicated in human disease for decades. Virtually all disease-causing aneuploidies arise as *de novo* events. The best-known example for this is trisomy 21, identified in 1959 as the cause of Down syndrome.[24] In the beginning of this millennium, genomic microarray technology provided insight into the role of *de novo* copy number variations (CNVs) in disease.[25] Even though large CNVs occur at a very low rate, arising at a frequency of only 0.01 to 0.02 events per generation,[25–27] they contribute significantly to severe and early-onset neurodevelopmental disorders and congenital malformations due to their disruptive effect on many genes.[28] The magnitude of the contribution of *de novo* genetic alterations to human disease, however, has only recently become fully apparent now that NGS approaches allow the reliable and affordable detection of all types of *de novo* mutations.[25] Damaging *de novo* point mutations and indels affecting important genes in development have been established as a prominent cause of both rare and common genetic disorders.[29–35]

In this chapter, we first touch on the biological aspects of *de novo* mutations in humans, such as their origin, distribution throughout the genome and factors related to their occurrence and timing. Later, we discuss the increasingly recognized role of *de novo* mutations in human disease and other translational aspects. Throughout this chapter, we focus mostly on *de novo* SNVs and refer to previous work from others for more information on the role of *de novo* CNVs and other structural genomic variation in human disease.[36,37]

Causes of *de novo* mutations

Mistakes during DNA replication can give rise to *de novo* mutations as a result of the erroneous incorporation of nucleotides by DNA polymerases.[38] DNA polymerases ε and δ catalyze replication predominantly in the leading and lagging strand, respectively. Both polymerases integrate nucleotides during polymerization in a highly selective way, with an average of one mismatch per 10^4-10^5 bp *in vitro*.[39,40] A proofreading subunit present in both polymerases

	Inherited Variants	*De novo* mutations
Single Nucleotide Variants (SNVs)	3.5 to 4.4 million[4]	44 to 82 [9,10,12,13,15]
Number of coding SNVs	22,186 [10]	1-2 [25]
Insertions and deletions (indels <50 bp)	~550,000 [4]	2.9 - 9 [26,81]
Large indels (50 – 5,000 bp)[a]	~1000 [4]	0.16 [26]
Copy Number Variations (CNVs)	~160 [4]	0.0154 [26, b]
Selection pressure in previous generation(s)	High	None
Damaging capacity of variants	Majority with small effect	High
Differences in population	Yes	None
Parental/Paternal age effect	None	Strong
Detection of variants	Imputable	Not imputable
Amenable to positional cloning[c]	Yes	No

Table 1. Comparison of inherited and *de novo* variants. [a] Due to technical limitations, the number and mutation rate for large indels ranging between 50 and 5000 bp remains uncertain. Novel sequencing approaches will likely provide more accurate estimates (see Chaisson *et al.*[237]). [b] per generation for CNVs larger than 100 kb. [c] Positional cloning by linkage analysis or homozygosity mapping.

subsequently verifies the geometry of the paired nucleotides to ensure that the incorporated base is correct.[38] Single or multiple base-pair mismatches may cause alterations in the structure of the replicating DNA and can be restored by the mismatch repair (MMR) pathway.[41] The MMR pathway is highly efficient, which explains why the amount of mutations generated during DNA replication is much lower than the polymerase error rate. The frequency at which specific base-pair substitutions arise may be different from the speed at which they are repaired, which defines the mutation rates for specific base-pair substitutions.[41] Incomplete repair can lead to single or multiple base-pair substitutions or indels. Additionally, damaged nucleotides can be incorporated during replication, leading to mispairings and base substitutions.[42]

DNA lesions may also appear spontaneously as a consequence of exogenous or endogenous mutagens; UV or ionizing radiation and DNA-reactive chemicals are an example of the former, while reactive oxygen species belong to the latter.[38] Prior to replication, these spontaneous lesions are repaired mainly by the nucleotide excision repair system and base excision repair pathways.[43] However, inefficient repair of pre-mutations before a new round of DNA replication may lead to the mutation becoming permanently fixed in either one or both daughter cells.[44] If mutation repair fails, DNA replication may also be completely arrested and ultimately lead to cell death.[44]

The difference between the rate at which pre-mutagenic damage appears in DNA and the rate at which it is repaired defines the rate at which *de novo* mutations arise. It is often assumed that germline *de novo* mutations originate from errors in DNA replication during gametogenesis, particularly in sperm cells and their precursors (see paragraph on parental origin of *de novo* mutations). However, inefficient repair of spontaneous DNA lesions may also give rise to *de novo* mutations during spermatogenesis, as continuous proliferation and short periods between cell divisions may translate into less time to repair these lesions.[44,45] In oogenesis, spontaneous DNA mutations coupled to inefficient repair mechanisms may play a more prominent role.[44] Therefore, while the *de novo* mutation rate is a reflection of the replication error rate and the number of mitoses a cell has undergone, this number is also influenced by the amount of time between mitoses and the efficiency of the DNA repair.[44]

Distribution of *de novo* mutations in the genome

While the average human mutation rate is 1-1.8 $\times$ 10^{-8} per nucleotide per generation,[1,9–13] mutagenesis does not occur completely at random across the genome.[9] Variation in mutability across different areas of the genome can be explained by intrinsic characteristics of the genomic region itself, related to its sequence composition and functional context.[46] Certain factors playing a role in the mutability of the genomic region are predicted to be shared by all cell types in the human organism. These include the local base pair context, recombination rate and replication timing.[9,13,47] Replication timing refers to the order in which different areas of the genome are replicated during the S-phase of the cell cycle. Genomic regions which are replicated late display more genetic variation than regions that are replicated early.[47] It has been suggested that this may be due to a higher mutability that is secondary to depletion of dNTP at the end of replication, although other changes such as alterations in polymerase activity and decreased MMR repair activity have also been implicated.[38,48,49]

Other factors influencing mutability may vary from cell to cell, depending on the transcriptional activity and chromatin state.[50–52] In addition, recent whole genome sequencing (WGS) studies have revealed the presence of so-called "mutational clusters" and "mutational hotspots". Mutational clusters correspond to the observation of multiple *de novo* mutations in very close vicinity in a single individual, while multiple *de novo* mutations occurring at the same location in several individuals are an indication of the existence of mutational hotspots.[53]

Nucleotide differences: transitions, transversions and CpGs

The molecular events underlying transitions occur more frequently than those leading to transversions, resulting in a two-fold rate of transitions over

transversions across the genome.[27,38] Transitions arise predominantly as a result of C>T mutations, which is at least partially explained by the mutability of CpG dinucleotides.[54] The cytosine in a CpG dinucleotide often undergoes methylation at the 5th position of the 6-atom ring leading to 5-methylcytosine (5-mC). In humans, methylated CpG dinucleotides are known to be chemically unstable and highly mutable due to deamination of 5-mC at CpG dinucleotides resulting in G:T mismatches.[12] Indeed, the mutability of CpG dinucleotides is approximately ten to eighteen times higher than that of other dinucleotides[27] and, as a result, CpG dinucleotides are found at only a fraction of their expected frequency in the human genome.[54] The high *de novo* mutation rate at CpG sites is also illustrated by the recent work of the Exome Aggregation Consortium (ExAC). Through the work of this consortium, exome data from more than 60,000 individuals without severe pediatric disease is currently available. Analysis of the data in ExAC shows that the discovery of new mutations at CpG dinucleotides reaches saturation at 20,000 exomes.[55,56] This emphasizes that identical CpG mutations do not necessarily reflect an ancestral event but are likely the result of independent *de novo* mutations.

Remarkably, the mutability of CpG dinucleotides is lower in genomic regions enriched for CpG and with higher GC content than in the rest of the genome.[44] In fact, the mutation rate for CpGs in the GC-richest regions of the genome are two to threefold lower than in the rest of the genome.[44,48] This could be the result of lower methylation levels, the effect of selection because the regions play a role in gene regulation or secondary to stronger binding between DNA strands impeding separation and spontaneous deamination.[38,44,57]

Mutational signatures underlying specific mutational processes

While errors in DNA replication, exposure to mutagens or failure to repair DNA damage can all result in mutations, there are differences in the pattern of mutations arising from each of these processes. A "mutational signature" has been defined as a pattern of mutations which is specific to a mutational process occurring in a cell, tissue or organism.[58] A recent study based on the analysis of 4.9 million somatic mutations in more than 12,000 cancer genomes defined 21 mutational signatures associated with mutational processes active in somatic cells (termed signature 1 to 21).[58] Detailed descriptions of each signature are available at http://cancer.sanger.ac.uk/cosmic/signatures. Each of these millions of mutations is placed into one of 96 possible mutation types based on six possible base pair substitutions (C>A, C>G, C>T, T>A, T>C and T>G) and one of four possible base pairs adjacent to the mutation both at the 5′ and at the 3′ position of the mutation. Concisely, each mutation type is a trinucleotide in which the middle base pair is mutated to a specific nucleotide and each mutational signature is defined by the frequency of each mutation type observed.[59]

A recent study showed that the mutational spectrum of germline *de novo* mutations correlated best with two of these previously described mutational signatures, currently known as signature 1 and 5.[11,13] This suggests that the mutational processes associated with these signatures in somatic cells may also be active in germ cells although the mechanisms underlying the processes remain elusive. Mutational signature 1 represents close to 25% of *de novo* germline mutations and is characterized by a high proportion of C>T transitions at CpG dinucleotides, which is associated with deamination of methylated cytosine.[11,58] Mutational signature 5, which corresponds to the remaining 75% of *de novo* mutations, is characterized mainly by A>G transitions.[11] While the mechanism underlying this signature remains unclear, the mutations observed as part of this signature may be secondary to spontaneous deamination of adenine to hypoxanthine, which is then read as guanine.[60] This mutational signature is associated with transcriptional strand bias, suggesting that some of these mutations arise from adducts subject to transcription-coupled repair.[60]

Mutational clusters and hotspots

De novo mutations occur throughout the human genome but occasionally several mutations may arise at a closer distance than expected by random distribution.[9] The term "mutational clusters" refers to the occurrence of *de novo* mutations in an individual at a closer distance than expected, with multiple *de novo* mutations within regions ranging from 10 to 100 kb.[9,12,13,53] Mutational clusters display a unique mutational spectrum with a lower rate of transitions and a large proportion of C>G transversions.[13] This phenomenon has been described to arise in somatic cells in the context of cancer, where it is known as "kataegis", and is linked to APOBEC enzymes.[53,58] It has been suggested that clusters involving C>G transversions could be related to the formation of single stranded DNA in diverse cellular processes, such as double-strand breaks and dysfunctional replication forks.[61] Single stranded DNA may be mistaken for retroelements and attacked by APOBEC enzymes, which convert cytosine to uracil.[53] The mutations are then repaired via base-excision repair and subsequent translesional DNA synthesis with error-prone polymerases.[38] Indeed, mutational clusters have been described to be reminiscent of APOBEC-mediated mutations albeit with a different sequence context.[12,13] The occurrence of mutational clusters has been found to correlate with increased parental age.[13]

Another origin for some of these clusters could be chromosomal re-arrangements. It has been shown that the mutation rate for SNVs is elevated and SNVs can cluster in proximity of the breakpoints of *de novo* CNVs.[62,63] This is likely the result of the replicative CNV mechanism in which a low-fidelity, error-prone DNA polymerase is used during repair of DNA. Indeed, work performed in yeast supports the observation that double-strand break-induced replication is a source of mutation clusters.[61]

In contrast to the mutation clusters that occur within one individual, mutational hotspots are considered overlapping loci that are found to be mutated more frequently than expected in different individuals. Recent research based on WGS datasets and modeling has identified such hotspots in coding sequences.[9] Furthermore, the existence of these mutational hotspots has been recently confirmed in a larger study which showed specific bins of 1 Mb within the human genome with elevated mutation rates.[13] Interestingly, in this study, two bins including genes *CSMD1* and *WWOX* were shown to have a higher maternal than paternal mutation rate. The mechanism for this is still largely unknown, but the latter is a well-known fragile site within the human genome.[64] Other sites of the human genome that are especially prone to *de novo* mutations include rDNA gene clusters,[65] segmental duplications[66] and microsatellites[67] with mutation rates three to four orders of magnitude higher than average.[68]

Timing of *de novo* mutations

De novo mutations occur predominantly in the egg or sperm cell resulting, upon fertilization, in a zygote with a constitutive germline mutation present in all of the cells of the organism. The advent of NGS allowed scientists to demonstrate that *de novo* mutations occur as non-germline events more often than previously estimated.[3,69–71] Mosaicism, which is the existence of two or more genetically distinct cell populations in an individual developing from a single fertilized egg,[72] is the norm rather than the exception. Postzygotic mutations, that is mutations arising in the first few cells divisions after fertilization, can lead to high level mosaicism and be present in many different tissues of an organism. Mutations that arise later in development or post-natal life, on the other hand, can remain restricted to a single tissue or even to a small number of somatic cells (see Figure 2). Although mutations which occur in somatic cells during post-natal and adult life are conventionally not considered *de novo* mutations, they are included in this section as they represent the end of the spectrum of timing in which novel mutations may arise.

De novo mutations arising during gametogenesis

In human embryos, the primordial germ cells (PGCs) emerge from the epiblast, 8 to 14 cell divisions after fertilization.[73] In these first cell divisions, the mutation rate appears to be similar in male and female embryos (approximately ~0.2-0.6 mutations per haploid genome per cell division, according to models estimating the mutation rate during gametogenesis).[11] After their specification, PGCs expand to form the pool of spermatogonial stem cells and the complete population of primary oocytes in male and female embryos, respectively.[11,73] Despite differences in the expansion of PGCs to oogonia or spermatogonia, the mutation rate during this step is similar in both sexes with approximately 0.5 to

0.7 mutations per haploid genome per cell division, according to computational modeling.[11] However, after puberty, the processes involved in spermatogenesis and oogenesis diverge further. Spermatogonial stem cells divide by mitosis approximately every 16 days, maintaining the spermatogonial stem cell pool while generating differentiated spermatogonial cells which produce sperm cells through an additional round of mitosis followed by meiosis.[74] In contrast, each menstrual cycle, a few oocytes escape from meiotic arrest and complete the first meiotic division. After ovulation, the oocyte becomes arrested once more until fertilization, when it completes the second meiotic division. Thus, after PGC expansion in embryogenesis, oocytes only undergo one additional round of DNA replication in their evolution to a mature ovum. In contrast, spermatogonial cells may undergo hundreds of rounds of DNA replication and cell division before their maturation to sperm cells.

Approximately 80% of all *de novo* germline point mutations arise on the paternal allele and advanced paternal age at conception has been established as the major factor linked to the increase in the number of *de novo* mutations in the offspring, both at the population level and within the same family (see Figure 2).[11,13,15] Spermatogonial cells continue to divide throughout life, which is likely to allow the progressive accumulation of mutations due to errors during DNA replication but also as a result of failure to repair non-replicative DNA damage between cell divisions.[44] Furthermore, the efficiency of endogenous defense systems against radical oxygen species and of DNA repair mechanisms may also decline with age.[75,76] *De novo* mutations in children of young fathers show a different signature and localize to later-replicating regions of the genome compared to children of old fathers, suggesting that additional factors contribute to *de novo* mutations with age.[12,13] It has been calculated that 1 to 3 *de novo* mutations are added to the germline mutational load of the offspring for each paternal year at conception, but this effect varies considerably between families.[11,13] This variability has been suggested to be due to individual differences in the rate of mutagenesis, in the frequency of spermatogonial stem cell division and even to genetic variation in DNA mismatch repair genes.[11] Indeed, one could speculate that deleterious variation in genes involved in replication and repair could predispose to elevated *de novo* mutation rates not only in somatic cells but also in the germline, as has been observed in mouse models lacking exonuclease activity in DNA polymerase δ.[77]

The effect of increased maternal age is well established for errors leading to chromosomal nondisjunction involved in aneuploidies,[78,79] but less so for *de novo* point mutations. The fixed number of mitoses required for oogenesis would entail that maternal age would not be linked to an increase in DNA replication-associated mutations. However, an effect of maternal age on the number of *de novo* mutations has been reported recently,[13,80] likely reflecting an excess of non-replicative DNA damage that is not properly repaired.[44] This maternal age effect

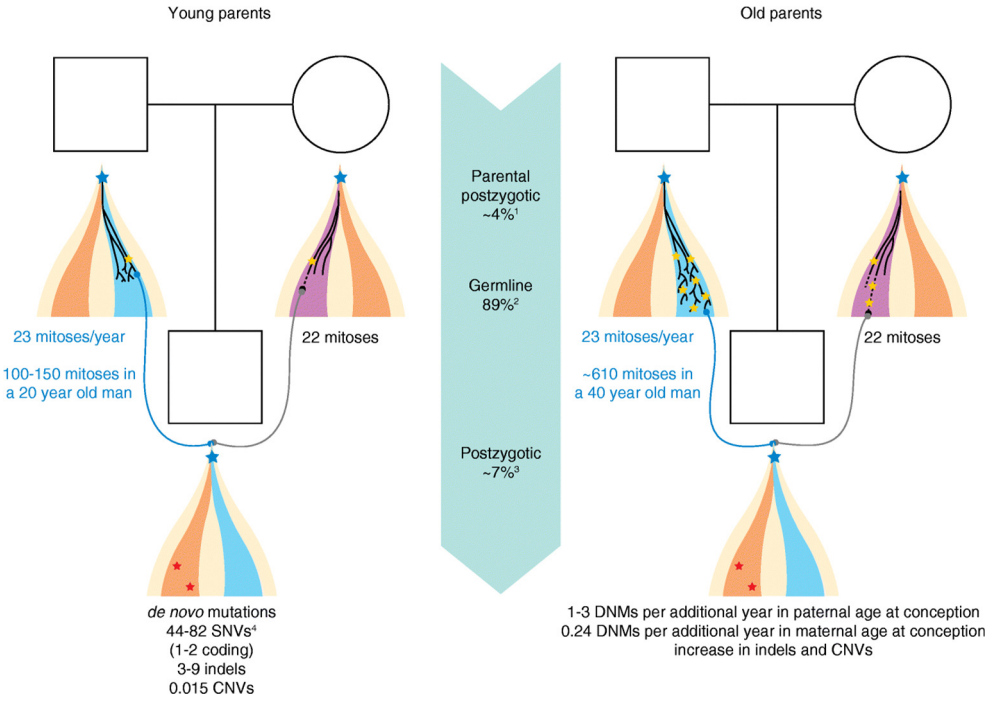


Figure 2. Timing of *de novo* mutations. Sperm cells have undergone approximately 100 to 150 mitoses in a 20-year-old man while oocytes have gone through 22 mitoses in a woman of the same age (left). As a result of errors in replication and repair of DNA damage occurring during parental embryogenesis, gametogenesis or as postzygotic events in the offspring, *de novo* mutations (DNMs) arise in each new generation. Advanced parental age is associated with an increase in the number of *de novo* mutations (right). The male germline adds 23 mitoses per year, entailing that a spermatogonial stem cell in a 40-year-old man has undergone more than 600 cell mitoses. Each additional year in paternal age at conception adds 1 to 3 *de novo* mutations to the genome of the offspring. Oogenesis has a fixed number of mitoses, but mutations accumulate over time possibly due to failure to repair DNA damage. The increase in number of *de novo* mutations with maternal age is lower: 0.24 extra *de novo* mutations for each additional year of maternal age at conception. Cell lineages modified from Scally.[236] Somatic cells are showed in orange, the male germline is shown in blue and the female germline is shown in purple. Blue stars represent postzygotic mutations present in the germline and in somatic cells; yellow stars represent mutations arising exclusively in the germline; red stars represent somatic mutations arising during embryonic development or post-natal life and are absent from germline cells. [1] The ratio of paternal to maternal mutations originating from parental gonosomal mosaicism is 1:1; [2] the ratio of paternal to maternal germline *de novo* mutations is 4:1; [3] the ratio of paternal to maternal postzygotic *de novo* mutations is 1:1; [4] this range is based on the average number of *de novo* mutations published in Michaelson *et al.*[9], Gilissen *et al.*[10], Francioli *et al.*[12], Goldmann *et al.*[13] and Kong *et al.*[15], regardless of parental age.

was initially reported in a study analyzing *de novo* mutations in WGS data from a large cohort of parent-offspring trios, in which maternal age correlated with the total number of *de novo* mutations after correcting for paternal age.[80] A more detailed analysis of the same cohort confirmed a subtle but significant increase in

the number of maternal *de novo* mutations with advancing maternal age, consisting of 0.24 additional *de novo* mutations per extra year of maternal age at conception.[13] Previous studies had failed to identify a maternal age effect on the number of *de novo* mutations.[12,15] This may be explained by differences in the parental age distribution between cohorts or due to lack of statistical power to detect this subtle effect for which paternal age is a confounder.[80] The increase of *de novo* mutations with advanced paternal and maternal age support the possibility that the accuracy of DNA repair mechanisms in germ cells decreases with age.[76]

De novo mutations arising postzygotically

Approximately 7% of seemingly *de novo* mutations are present in blood as high level mosaic mutations, having likely occurred as early postzygotic events.[70,71,81] Considering an average of 44 to 82 *de novo* mutations in our genome,[9,10,12,13,15] it is likely that each of us carries between 8 and 15 early postzygotic *de novo* mutations in our organism. This, together with the observation that chromosomal instability and structural rearrangements are common in cleavage-stage human embryos, has led to the suggestion that early embryogenesis may be a period of high mutability.[82,83] Prior to the initiation of transcription and translation in the zygote, human embryos rely on maternal proteins contributed by the oocyte,[84] which could lead to shortage of proteins involved in DNA replication and repair, resulting in genomic instability.[3]

Recent studies in zebrafish have shown that the majority of cells in each organ derive from a small number of progenitor cells,[85] while research in mouse embryos suggests that an individual progenitor cell can contribute unequally to multiple tissues.[86] Therefore, depending on the timing at which a *de novo* mutation arises during embryonic development, it may be present at different levels in multiple tissues or can be organ-specific.[87] A recent study examined multiple samples from the same individual and showed the widespread presence of postzygotic *de novo* mutations in tissues of different embryonic origin, including somatic and germ cells.[88] Furthermore, mutations may arise in the germ cell lineage after the specification of PGCs during early embryonic development, remaining isolated from somatic cells.[3] Although these mutations are undetectable in sampled tissues such as blood or buccal swabs, they can be transmitted to the offspring as germline events.

Somatic mutations arising during postnatal and adult life

Studies in mouse models have shown that the mutation frequency is higher in somatic cells than in germ cells,[89,90] a finding supported by the comparison of the somatic and germline mutation rate in humans.[76] Indeed,

somatic cells have been shown to accumulate hundreds of mutations throughout post-natal and adult life.[91] Large chromosomal abnormalities have been observed in many tissues in the human body[92] such as the blood, where the presence of these lesions increases with age.[93–95] For instance, loss of chromosome Y in blood cells has been described as a frequent event in aging males, affecting over 15% of men of 70 years of age or older.[96,97] Somatic mutations resulting in low level mosaicism are prevalent in healthy tissues,[98] including the brain,[99] blood[100–102] and skin, where the somatic mutation burden has been calculated at two to six SNVs per Mb of coding sequence per cell.[103]

The increased mutation frequency observed in somatic tissues may partially be explained by differences in the efficiency of DNA replication and repair mechanisms in germ and somatic cells.[76] However, the number and type of mutations observed in different somatic tissues in the human organism is highly variable.[104] This is likely due to disparities in the type of environmental factors that cells are exposed to,[58] as well as the number of stem cell divisions per tissue.[105] Indeed, exogenous factors such as radical oxygen species related to diet and/or lifestyle have been shown to be a major source of DNA damage in adult tissues.[106] Different tissues are subjected to different mutagens. For instance, skin is exposed to ultraviolet light, which contributes substantially to the mutations observed in this tissue,[103] while factors such as tobacco and alcohol consumption are important sources of mutations in the esophagus.[107]

As for differences in the number of stem cell divisions per tissue, a careful balance between cell proliferation, differentiation and apoptosis is necessary for the maintenance of developed tissues in adult organisms. Adult stem cells are multipotent cells that divide to sustain their own population and differentiate to generate specialized cells which usually form the bulk of adult tissues. Differentiated cells undergo different degrees of turnover; some tissues in the human organism undergo constant renewal, such as the lining of the gut,[108] while others, like the frontal lobe of the brain, do not go through any turnover. Stem cells in tissues with constant renewal undergo more cell divisions than those of tissues that do not[105] and are therefore more likely to acquire mutations associated to DNA replication. Because stem cells are long-lived, mutations arising in stem cells can persist in the tissue for years and accumulate dozens of mutations over the course of time.

De novo mutations in human disease

The effects of *de novo* mutations encompass a wide spectrum, ranging from beneficial to an organism to lethal. Within this phenotypic spectrum, *de novo* mutations can cause many types of human disease including severe congenital disorders and late-onset diseases. The medical relevance of *de novo* mutations has only recently been fully appreciated, mainly because advances in

sequencing technology have allowed a comprehensive analysis of these mutations.[25] NGS techniques such as whole exome sequencing (WES) or WGS now provide the possibility to detect most, if not all, genetic variation present in a patient. To this end, trio-based WES or WGS has been instrumental in detecting and characterizing *de novo* mutations in patients with a wide variety of diseases.[25,35]

De novo mutations in pediatric disease

De novo mutations are now well known to play an important role in severe early-onset diseases, which for the most part arise sporadically because of their impact on fitness; due to the severity of the phenotype in which they often result, an individual with a deleterious *de novo* mutation will not produce offspring and the phenotype therefore only arises through *de novo* mutations.

In the first 5 years of widespread availability of WES, more than 500 novel disease-gene associations have been identified, with the strongest increase in sporadic diseases caused by *de novo* mutations.[35,109,110] Recent studies applying exome sequencing in the clinic have shown that of all sporadic cases who received a molecular diagnosis through clinical exome sequencing, between 60% and 75% could be explained by *de novo* mutations.[111,112] *De novo* mutations affecting the coding region have also been established as an important cause of common neurodevelopmental disorders such as autism,[29,30] epilepsy[31] and intellectual disability,[33,34] which affect over 1% of the population.[113,114] Clearly, these common genetic disorders are not explained by *de novo* mutations affecting the same locus in every patient. Instead, an extreme genetic heterogeneity is observed and patients with common genetic disorders carry *de novo* mutations in many different genes. The population frequency of a disorder caused by *de novo* mutations is determined in large part by the number of genes or genetic loci that can result in this disorder when mutated, which we have referred to previously as the mutational target.[25] Rare disorders are most often caused by mutations in a single gene or a small number of genes, while common genetic disorders usually have a large mutational target often consisting of hundreds to thousands of genes or genetic loci.[25] As an example, more than 700 genes have now been identified to cause autosomal dominant intellectual disability when mutated,[113] and this number is rapidly increasing since the widespread application of NGS technology. Based on these sequencing studies, it appears that the majority of the most severe neurodevelopmental phenotypes, such as severe intellectual disability with an IQ below 50, are the consequence of damaging *de novo* germline mutations in the coding region.[10] An enrichment for damaging *de novo* mutations has also been observed in individuals with milder phenotypes such as autism spectrum disorder without cognitive deficits.[16,18,29,30,115] For these milder phenotypes which have less impact on fitness, the exact contribution of *de novo* mutations to the disease burden is not yet

firmly established and inherited variation is likely to be at least as important in the expression of the phenotype.[116–118] Next to neurodevelopmental disorders, *de novo* mutations also play a prominent role in pediatric diseases such as congenital heart defects (CHD).[119–121] In agreement with the observation made in neurodevelopmental disorders, recent studies found the highest contribution of *de novo* mutations to disease in individuals with the most severe and syndromic forms of CHD.[119,121] Finally, it is essential in large-scale sequencing studies to formally test whether the recurrence of *de novo* mutations in a gene exceeds the number of observations expected by chance.[122]

The vast majority of pathogenic *de novo* mutations are involved in dominant genetic disorders. This appears logical, as a single damaging *de novo* mutation can be sufficient to cause these kinds of disorders. However, there are examples of recessive disorders which can be caused by the combination of an inherited mutation on one allele and the occurrence of a *de novo* mutation on the other.[33] In a cohort of 100 trios with severe ID, we identified one case of autosomal recessive intellectual disability due to the inheritance of one pathogenic allele and a *de novo* hit in the other[33] and similar observations in the context of late-onset disease are described below. Furthermore, there are reports of cases with a merged phenotype composed of two clinically distinct disorders of which either one or both are caused by a pathogenic *de novo* mutation.[111] Phenotype-based and classic genetic approaches are insufficient to diagnose individuals with this kind of combined disease, illustrating the power of an unbiased genotype-first approach. Additionally, this approach reduces the need for clinical homogeneity for disease-gene identification studies as was required for phenotype-first approaches.[123,124]

De novo mutations in late-onset disorders

Few studies until now have addressed the role of *de novo* mutations in late-onset diseases. The role of *de novo* mutations is likely to be smaller in late-onset disorders than in pediatric disorders given the effect of *de novo* mutations on reproductive fitness. Still, genes involved in adult-onset disorders are just as likely to be affected by *de novo* mutations as genes involved in pediatric disorders. A complicating factor in these late-onset disorders, however, is the collection of parental samples for the study of *de novo* mutations.[125] Despite this obstacle, recent publications have suggested a link between *de novo* mutations and late-onset neurological and psychiatric disorders: Parkinson's disease, amyotrophic lateral sclerosis, schizophrenia and bipolar disorder have been associated with *de novo* SNVs and CNVs.[126–133] For example, one study found that 10% of individuals with sporadic schizophrenia have a rare *de novo* CNV compared to 1.26% of controls.[128] Exome sequencing of a cohort of 623 schizophrenia trios identified an enrichment for *de novo* point mutations in genes encoding synaptic proteins in cases compared to controls.[126] A large meta-

analysis recently identified both an excess of loss-of-function mutations in *SETD1A* and an excess of *de novo* occurrence of these mutations in individuals with schizophrenia compared to controls.[134] Recent studies have exposed a genetic overlap between neurodevelopmental disorders and schizophrenia, with *de novo* mutations in the same gene being involved in both early and late-onset disorders.[134–136] While *de novo* mutations have been firmly linked to neurodevelopmental disorders, their involvement in late-onset psychiatric phenotypes is more controversial. This may be the result of a more complex underlying genetic architecture,[137] together with a more prominent role for environmental factors in the expression of the phenotype.[138] Furthermore, based on the high degree of mosaicism observed in a normal human brain, it has been suggested that that pathogenic postzygotic and somatic mutations could contribute to psychiatric disorders.[139,140]

Cancer, particularly in relatively young individuals without relevant family history, has been associated with *de novo* mutations in genes involved in cancer predisposition syndromes. For example, at least 7% of germline mutations in *TP53* in individuals with Li-Fraumeni syndrome occurred *de novo,*[141] and a similar proportion has been identified for mutations in *APC* involved in familial adenomatous polyposis.[142] Nevertheless, the rate of *de novo* mutations in genes involved in other cancer predisposition syndromes, such as *BRCA1* and *BRCA2*[143] or in DNA mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*)[144] has been reported to be much lower. Postzygotic *de novo* mutations have been implicated in early-onset cancer[145,146] and could well represent an early mutational event in the development of cancer in the general population.[147]

Interestingly, *de novo* mutations have also been identified as causative mutations in genetic disorders which are typically inherited, such as hereditary blindness. For instance, the rate of causative *de novo* mutations among sporadic cases within a cohort of patients with Retinitis Pigmentosa was close to 10%,[148] a result which was later confirmed by an independent study.[149] Although for the majority of this group, the *de novo* mutation represented a single dominant hit causative for the phenotype, in one case the *de novo* mutation was in fact the second hit in an autosomal recessive form of Retinitis Pigmentosa. Similarly, in a cohort suffering from mild to moderate sensorineural hearing loss, *de novo* mutations were identified in 2 out of 11 sporadic cases,[150] also suggesting a role for *de novo* mutations in this heterogeneous disorder.

As *de novo* mutations are known to play an important role in disorders that affect fitness, it may also be very relevant to investigate their role in disorders linked to fertility, such as male infertility. Both *de novo* chromosome Y deletions as well as *de novo* point mutations in a few genes have been found to cause this disorder,[151,152] but a systematic screen is lacking so far.

The role of timing of *de novo* mutations in human disease

As a consequence of technological improvements allowing the detection of (low level) mosaic mutations at a genome-wide scale, mosaicism has been recognized to have an important role in human biology and disease. Postzygotic and somatic mutations have been recently implicated in several human diseases, ranging from developmental disorders[153–155] to cancer.[145–147] A lesson derived from these discoveries is that the timing of a pathogenic *de novo* mutation can have an important influence on the expression of the phenotype.

Mutations arising during gametogenesis

Pathogenic *de novo* mutations arising during gametogenesis cause a clinical phenotype in the offspring in whom they are present as a constitutive event. However, the consequences of these mutations may not be limited to the next generation, but could also affect the cells in which they arise. Thus, certain *de novo* mutations arising during gametogenesis may also have an effect on cellular behavior in spermatogonial stem cells or oogonia in the parent of origin for the mutation.

A small subset of *de novo* mutations which are highly recurrent and localize to specific nucleotides in the genome have been observed to have a striking increase with paternal age. These *de novo* mutations are thought to grant spermatogonial stem cells a growth advantage, leading to clonal expansion of mutated cells in the testis.[156] For instance, gain-of-function mutations in genes in the RAS/MAPK pathway have been shown to cause clonal expansion of mutant spermatogonial stem cells due to proliferative selective advantage.[156,157] Computational modeling suggests that this would result from a slightly increased ratio of symmetric versus asymmetric divisions in mutant spermatogonial stem cells, favoring the production of two mutated spermatogonial stem cells compared to a single mutated stem cell and one differentiated spermatogonial stem cell harboring the mutation.[158,159] Therefore, over time, spermatogonial stem cells carrying these mutations undergo positive selection due to higher self-renewal than surrounding wild-type cells and expand clonally in the testis.[160] The occurrence and enrichment of mutations in spermatogonial stem cells is thought to take place in all men and would entail that the testes of older men contain a higher number of clones of mutant spermatogonial stem cells.[156,157]

Interestingly, the first mutations implicated in clonal expansion in spermatogonial stem cells were initially shown to cause developmental disorders such as Noonan and Costello syndrome (caused by *PTPN11* and *HRAS* mutations, respectively),[157,160,161] Apert, Crouzon and Pfeiffer syndromes (*FGFR2*),[160,162] achondroplasia, Muenke syndrome and thanatophoric dysplasia (*FGFR3*)[160,161] and multiple endocrine neoplasia (*RET*).[163] Mutations which are positively selected at the spermatogonial stem cell level but are detrimental at the organism level have

been termed to behave selfishly and are therefore referred to as "selfish mutations".[161] Due to the expansion of mutant cells over time, the incidence of these developmental disorders shows an exponential increase with paternal age at conception, well beyond the increase observed for other disorders caused by *de novo* mutations.[164] Appropriately, these disorders are known as recurrent, autosomal dominant, male-biased and paternal age effect disorders (RAMP) or, simply, paternal age effect disorders (PAE).[45,157] Because of the selfish selection of mutant spermatogonial cells, PAE disorders have an incidence up to 1000-fold higher than expected based on the mutational target size and the average mutation rate.[45,164] It has been hypothesized that "selfish mutations" with a weaker effect on spermatogonial stem cell behavior could be involved in more common phenotypes such as intellectual disability, autism or epilepsy.[165] Furthermore, "selfish" behavior is a characteristic of certain mutations driving cancer, as they lead to positive cellular selection despite being harmful for the organism. Predictably, several mutations behaving selfishly in spermatogonial stem cells have also been identified as somatic events driving clonal growth in tumorigenesis.[161]

Following the identification of genomic regions enriched for maternal *de novo* mutations,[13] the possibility of selfish mutations in the maternal germ line has also been put forward.[76] It appears these genomic regions harbor genes with a role in tumor suppression and some *de novo* mutations could speculatively provide mutant oocytes in aging women with a survival advantage over wild-type ones.[76]

Postzygotic mutations

Regardless of whether they occur in the germline or postzygotically, some *de novo* mutations lead to a single Mendelian phenotype in which the mosaic and constitutive form are part of the same clinical spectrum.[166] For example, pathogenic mutations in genes involved in epileptic encephalopathies[167] and cerebral cortical malformations[168] have been shown to cause similar phenotypes when they arise either in the germline or as postzygotic *de novo* mutations leading to mosaicism in the brain. However, in some of these cases, mosaicism may cause a milder clinical phenotype than a constitutive mutation.[169,170]

De novo mutations can also result in different phenotypes when they are present in the germline or arise postzygotically.[171] Some *de novo* mutations lead to developmental disorders only if the *de novo* mutation occurs postzygotically, as the constitutive presence of the mutation is suspected to be lethal.[172,173] Examples of this include Proteus syndrome (caused by *AKT1* mutations),[153] Sturge-Weber syndrome (*GNAQ*)[154] and CLOVES syndrome (*PIK3CA*).[174] A common feature to these disorders is that they are caused by mutations known to lead to activation of cellular proliferation pathways and overgrowth. The

mutations with the strongest effect generally result in more severe developmental alterations,[175] suggesting that the type of *de novo* mutation influences the expression of the phenotype. Remarkably, the mutations with the strongest effect on activation have also been observed as somatic events in cancer,[175] for which constitutive activation of cellular proliferation pathways is a major hallmark.[176] This finding supports that not only the type, but also the time at which a pathogenic mutation occurs is crucial in defining its consequences.

The timing of a postzygotic mutation determines the percentage of affected cells in the organism and the type of tissues involved.[72,154] For instance, the same genetic alteration in genes in the RAS/MAPK pathway can result in very diverse phenotypes depending on the timing at which they arise.[171,177,178] Mutations in *HRAS* mutating codon G12 of the HRAS protein have been identified in Costello syndrome when present in the germline,[179] but postzygotic and embryonic occurrence of mutations in this residue has been observed in Schimmelpenning syndrome,[171] sebaceous nevus,[171] keratinocytic epidermal nevi[180] and in early-onset bladder cancer[145,181]. Furthermore, identical mutations in *PIK3CA* can cause different phenotypes ranging from different overgrowth syndromes[155] to lymphatic[182] and venous malformations[183] depending on the tissue distribution. Therefore, the timing of a pathogenic *de novo* mutation is likely instrumental in defining its phenotypic consequences as it determines the burden placed by the mutation upon the organism, including the type of tissues affected and the percentage of cells in which the mutation is present.[72,154]

An important characteristic of postzygotic mutations is that they generate genetically distinct populations of cells that coevolve within a single organism. As such, mosaicism may lead to differences in cell fitness which could result in competition between the mutant and wild-type cells in the developing human embryo. Indirect or classic cell competition is the context-dependent elimination of cells with reduced metabolism in a developing embryo; "fit cells" induce apoptosis of cells that are less fit but otherwise viable and proliferate to occupy their place.[184] Classic cell competition was originally described in *Drosophila* as a mechanism to eliminate suboptimal cells during embryogenesis and ensure fitness of the cells from which the organs and the multicellular organism would develop.[185] Classic cell competition was recently shown to also take place during early mammalian embryogenesis.[186,187]

Another possible consequence of postzygotic *de novo* mutations is that genetically different cells in a developing organism may fail to coordinate due to heterogeneous responses to stimuli. This phenomenon was first observed in the context of mutations in *EFNB1* in craniofrontonasal syndrome and *PCDH19* in epileptic encephalopathy.[188–190] The proteins encoded by *EFNB1* and *PCDH19* play a role in forming topologically-defined tissue domains. Consequently, cellular mosaicism for these proteins leads to abnormalities in cellular determination and migration, alterations in the formation of compartments and mosaic tissue segments. In both conditions, females present a severe phenotype

while male carriers are asymptomatic or mildly affected. This is explained by random inactivation of the X chromosome in females leading to mosaicism in the cellular expression of the genetic mutation, causing discordant cellular responses and resulting in alterations in tissue organization. On the other hand, cells of hemizygous carrier males express the genetic mutation uniformly and compensate through the use of highly homologous receptors which allows signaling to bypass the mutated protein. Remarkably, mosaicism for pathogenic mutations in these genes has been identified in clinically affected males, for whom the molecular mechanism of disease is similar as for females with random X inactivation.

Finally, pathogenic *de novo* mutations arising postzygotically and leading to gonadal or gonosomal mosaicism may be clinically silent in an individual. Nevertheless, this person has increased risk of transmitting this pathogenic mutation to the next generation as a constitutive event, resulting in a clinical disorder in his or her offspring.[191]

Somatic mutations

Pathogenic mutations arising somatically can have an effect on cell behavior. Depending on their effect, these mutations can promote their own disappearance or propagation; while cells acquiring deleterious mutations are likely to die, some mutations may grant cells an advantage and promote their survival or replication.[192] Clonal expansion is a phenomenon by which a single cell forms groups of identical daughter cells. Clonal expansion can be the result of a stem cell favoring proliferation over its differentiation, as is thought to occur with spermatogonial stem cells harboring 'selfish' mutations. Another mechanism by which clonal expansion may occur is by escaping quiescence or having increased survival in comparison to other cells, as in revertant mosaicism in individuals with genetic diseases. In these individuals, clones of wild-type cells are occasionally observed in tissues such as skin or blood, due to correction of the deleterious alleles by spontaneous mutation and subsequent selection for wild-type cells in the context of pathogenic cells.[193–196]

Certain recurrent mutations arising somatically have been shown to be implicated in the clonal expansion of mutated hematopoietic stem cells detectable in circulating blood cells or "clonal hematopoiesis". Somatic mutations involved in clonal hematopoiesis most commonly affect *DNMT3A*, *TET2* and *ASXL1.*[101] These somatic mutations are found in blood-derived DNA at allelic fractions ranging from 2 to 10%, suggesting thereby that between 4 and 20% of nucleated cells in blood derive from mutated hematopoietic stem cells.[102] Previous work has shown that *DNMT3A* and *TET2* mutations involved in clonal hematopoiesis impair the differentiation of hematopoietic stem cells and increase self-renewal capacity, leading to a clonal expansion of these cells in the bone

marrow.[100,197] Loss of *ASXL1* leads to aberrant gene expression and epigenetic alterations secondary to loss of gene repression.[198] Clonal hematopoiesis is rarely observed in individuals younger than 50 years but increases in frequency with age, affecting up to 10% of individuals older than 65.[100,102,199] It has been suggested that the aging cellular background plays an important role in the selection and expansion of hematopoietic stem cells carrying mutations.[200] Aging leads to a decline in the function of hematopoietic stem cells[199] and myeloid bias[201] as well as modifications in the bone marrow niche.[202] It is therefore possible that certain mutations provide a cellular advantage in the context of the aging bone marrow, allowing for clonal expansion.[200]

Most mutations involved in clonal expansion have been identified with several types of tumors and are classified as oncogenic.[45] Indeed, the selective advantage granted to mutant stem cells and resulting in clonal expansion could contribute to the formation of tumors: spermatocytic seminomas in the testicle, in the case of spermatogonial stem cells[45] or to leukemia, in the case of hematopoietic stem cells.[100,203] However, these mutations are not sufficient to cause cancer; the transformation of a normal somatic cell into a cancer cell has been suggested to require at least six genetic mutations (granting self-sufficiency for growth signals, insensitivity to growth-arresting signals, evasion of apoptosis, ability to replicate without limits, sustained angiogenesis and tissue invasion and metastasis).[202] Adult stem cells have a long life span, which allows them to progressively accumulate the necessary genetic mutations to pave the way for tumorigenesis.[202] However, the acquisition of mutations is not linear, as cells acquiring advantageous mutations can also be eliminated due to stochasticity within the organism.[204,205]

De novo mutations in clinical practice

The recent recognition of the importance of *de novo* mutations in human disease has many implications on routine genetic testing and clinical practice. Germline *de novo* mutations are now established as the cause of disease in a large fraction of patients with severe early-onset disorders ranging from rare congenital malformation syndromes[206,207] to more common neurodevelopmental disorders, such as severe forms of intellectual disability,[33] epilepsy[31] and autism.[29] Together, these disorders represent a substantial proportion of all patients seen at neuropediatric and clinical genetics departments around the world.

Pinpointing the genetic cause of a disorder caused by a *de novo* mutation in an individual can be challenging from the clinical point of view because of pleiotropy as well as genetic heterogeneity underlying a single phenotype. For instance, intellectual disability can be caused by *de novo* point mutations, indels or CNVs in any of hundreds of genes.[113] This obstacle to providing a clinical diagnosis strongly argues for a reliable and affordable genomics approach which

can be used to detect these *de novo* mutations in large groups of patients. Exome and genome sequencing (which additionally offers the possibility of accurate detection of structural variation) of patient-parent trios is ideal for this and will soon become the first-tier diagnostic approach for these disorders. A key advantage of this trio-based sequencing approach is that it helps prioritize candidates by *de novo* occurrence, allowing clinical laboratories to focus on the most likely candidate mutations for follow-up and interpretation.[208]

Although identifying *de novo* mutations is becoming increasingly easy, interpreting them (*i.e.* linking them to a phenotype) often remains challenging.[209] Clinical interpretation of *de novo* mutations requires evaluation at the level of the affected locus or gene, as well as at the variant level.[210] The interpretation of candidate *de novo* mutations can be guided by the use of *in silico* prediction programs like SIFT, PolyPhen, MutationTaster and CADD.[210–213] Large scale databases of genetic variation can be used to see whether a gene or gene region shows constraint against variation in controls, as the frequency of a mutation in the population is often a good indirect estimation of its pathogenicity.[212] To this end, RVIS and selective constraint scores have become routine in the interpretation of *de novo* variants both in research and in the clinic.[122,214] Population databases such as ExAC are expected to be depleted of *de novo* disease-causing mutations for severe and early-onset disorders.[55] Given that *de novo* mutations are the rarest type of variation, the absence of a mutation from the ExAC database is not in itself sufficient evidence for its pathogenicity. On the other hand, the presence of a mutation in ExAC does not automatically entail that the mutation is not disease-causing. Pathogenic mutations involved in dominant disease are present in ExAC,[55] which may be explained by variable penetrance for these variants,[215] the presence of false-positive variants in the control database,[216] undiagnosed disease in control individuals or late-onset of disease. Possible other explanations for these observations could be the presence of these mutations as somatic events in control individuals[100–102] or resilience to disease in few selected individuals.[217]

In the context of research, evidence linking a gene or a mutation to a phenotype is traditionally established experimentally,[210,212] although functional validation is laborious and the necessary assays may differ per gene and per mutation. Many recent developments can support the interpretation of *de novo* mutations in human disease. For instance, to study the consequences of a mutation, induced pluripotent stem cells from patient-derived samples can be differentiated into cell types relevant for the respective disease.[218] Furthermore, as a robust method for *in vitro* and *in vivo* genetic manipulation, CRISPR/Cas9 can be used to establish cell and animal models for functional studies.[219,220] Other CRISPR/Cas9-based methods, such as Saturation Genome Editing, hold promise for the evaluation of hundreds of mutations in a single assay,[221] allowing the interpretation of *de novo* mutations to keep pace with their discovery in the genomics era.

The impact of mosaicism in human disease is not yet fully appreciated, possibly because postzygotic mutations often remain undetected. In comparison with germline *de novo* mutations, postzygotic mutations are more challenging to identify as they are only present in a small percentage of cells and may resemble false-positive artefacts in sequencing data. The detection of postzygotic mutations can be improved by using higher sequencing coverage and implementing technological improvements such as single molecule tracing, which is based on the incorporation of unique molecular identifiers to each DNA molecule during capture.[70,222] It is to be expected that our understanding of postzygotic mutations will increase rapidly in the coming years due to this and other improvements, as well as access to DNA from other (affected) tissues. Sequencing of cell-free DNA in plasma is currently being explored as a source of DNA from multiple tissues.[223–225]

The identification of a *de novo* mutation as the cause of disease in a patient has several implications for the patient and his or her family. First, the detection of the genetic defect underlying the phenotype establishes a genetic diagnosis which can be used to provide a prognosis based on data from other patients with similar mutations,[226] information about current treatment options[227] and, in the future, to the development and application of personalized therapeutic interventions.[228] Furthermore, the identification of a *de novo* mutation offers the parents of the affected patient an explanation as to why the disorder occurred and may help deal with feelings of guilt.[229,230] In terms of family planning, the identification of a *de novo* mutation as the cause of disease in a child may be positive news with regard to recurrence risk, as it is much lower than for recessive or dominant inherited disorders (slightly above 1% versus 25 and 50%, respectively).[11,191] However, the recurrence risk is strongly dependent on the timing of the mutation as parental mosaicism for the mutation increases the risk of recurrence.[191] Approximately 4% of seemingly *de novo* mutations originate from parental mosaicism detectable in blood[11] and recent work suggests that transmission of parental mosaicism could explain up to 10% of *de novo* mutations in autism spectrum disorder.[231] This entails that a fraction of *de novo* mutations has an estimated recurrence risk above 5%.[191] Furthermore, close to 7% of seemingly *de novo* mutations arise as postzygotic events in the offspring.[70,71,81] Parents of an individual with a postzygotic mutation have a low risk for recurrence of the mutation in an additional child, estimated to the same as the population risk.[72] Targeted deep sequencing of a disease-causing mutation can be performed to test for its presence in parental blood and detect mosaicism in the offspring. Although it is not yet offered on a routine basis, this kind of testing can provide a personalized and stratified estimate of the recurrence risk based on the presence or absence of mosaicism in the parents or in the offspring.

Finally, it is impossible to prevent *de novo* mutations from arising in the germline of each new generation but attention must be brought to the factors that increase the number of *de novo* mutations in the offspring. The single most

important risk factor is advanced paternal age at conception,[15] which is of great importance from an epidemiological perspective since most couples in Western countries are having children at later ages. In fact, this increase in *de novo* mutations with paternal age at conception may explain epidemiological studies that link increased paternal age to increased risk of neurodevelopmental disorders in offspring.[232] A recent population genetic modeling study, however, indicated that *de novo* mutations may not explain much of the increased risk of psychiatric disorders in children born to older fathers.[118] While this may be the case for relatively mild and late-onset phenotypes such as schizophrenia, *de novo* mutations are responsible for the majority of the most severe pediatric disorders arising in outbred populations.[10,233] At present, most attention, advice and guidelines are focused on advanced maternal age as a public health issue. It is evident from current work on *de novo* mutations that advising the public, including policy makers, on potential risks of advanced paternal age and the burden it may bring on society, is crucial. An extreme solution if reproduction is to be postponed may be to promote cryopreservation of oocytes and sperm,[234] a measure under much debate that has been termed "social freezing".

Conclusions and future directions

Advances in sequencing technologies have provided us with the ability to systematically identify most if not all *de novo* mutations in a genome. This has boosted fundamental research into the evolution of our genome by providing insight into the mechanisms that play a role in mutagenesis, the origins of these mutations and their distribution throughout the genome. While most of this research has been focused on germline mutations, we now see a shift towards the detection and study of somatic *de novo* mutations also for non-cancer phenotypes, greatly facilitated by more accurate and deeper coverage sequencing technologies. Next generation sequencing has also boosted research and diagnostics on sporadic diseases. The routine detection of *de novo* mutations by trio-based sequencing of patients and their unaffected parents in research as well as in diagnostics will soon allow the identification of most disease-causing genes involved in sporadic monogenic disorders. This will allow for the classification of different developmental and neurodevelopmental disorders based on the underlying genotype rather than solely on the phenotype. In turn, this offers the possibility of targeted medical consultations and interventions, engagement in gene-specific patient groups and, in some cases, in treatment. The study of *de novo* mutations will shift more and more towards the detection and characterization of non-coding *de novo* mutations in disease. Although a phenomenal challenge that will require large study cohorts and detailed functional validation, the limited number of *de novo* mutations per genome reduces the search space for pathogenic non-coding mutations, as was recently shown for non-coding de novo CNVs.[235]

References

1. Roach, J. C. *et al.* Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328,** 636–639 (2010).
2. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci **107**, 961–968 (2010).
3. Campbell, I. M., Shaw, C. A., Stankiewicz, P. & Lupski, J. R. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet **31**, 382–392 (2015).
4. Auton, A. et al. A global reference for human genetic variation. Nature **526**, 68–74 (2015).
5. Lupski, J. R., Belmont, J. W., Boerwinkle, E. & Gibbs, R. A. Clan Genomics and the Complex Architecture of Human Disease. Cell **147**, 32–43 (2011).
6. Haldane, J. B. S. The rate of spontaneous mutation of a human gene. J Genet **31**, 317–326 (1935).
7. Nachman, M. W. Haldane and the first estimates of the human mutation rate. J Genet **87**, 317–317 (2008).
8. Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. Hum Mutat **21**, 12–27 (2003).
9. Michaelson, J. J. et al. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell **151**, 1431–1442 (2012).
10. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. Nature **511**, 344–347 (2014).
11. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. Nat Genet **48**, 126–133 (2015).
12. Francioli, L. C. et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet **47**, 822–826 (2015).
13. Goldmann, J. M. et al. Parent-of-origin-specific signatures of de novo mutations. Nat Genet **48**, 935–939 (2016).
14. Conrad, D. F. et al. Variation in genome-wide mutation rates within and between human families. Nat Genet **43**, 712–714 (2011).
15. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature **488**, 471–475 (2012).
16. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature **485**, 246–250 (2012).
17. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature **485**, 237–241 (2012).
18. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature **485**, 242–245 (2012).
19. Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science **349**, 1343–1347 (2015).
20. Yi, X. et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science **329**, 75–78 (2010).
21. Bersaglieri, T. et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am J Hum Genet **74**, 1111–1120 (2004).
22. Fu, W. & Akey, J. M. Selection and Adaptation in the Human Genome. Annu Rev Genomics Hum Genet **14**, 467–489 (2013).
23. Hurst, L. D. Fundamental concepts in genetics: Genetics and the understanding of selection. Nat Rev Genet **10**, 83–93 (2009).
24. Lejeune, J., Gautier, M. & Turpin, R. Etude des chromosomes somatiques de neuf enfants mongoliens. [Study of somatic chromosomes from 9 mongoloid children]. Comptes rendus Hebd des se• ances l'Acade• mie des Sci **248**, 1721–2 (1959).
25. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. Nat Rev Genet **13**, 565–575 (2012).
26. Kloosterman, W. P. et al. Characteristics of de novo structural changes in the human genome. Genome Res **25**, 792–801 (2015).

27. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. Trends Genet **29**, 575–584 (2013).
28. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet **14**, 125–138 (2013).
29. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature **515**, 216–221 (2014).
30. O'Roak, B. J. et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet **43**, 585–589 (2011).
31. Allen, A. S. et al. De novo mutations in epileptic encephalopathies. Nature **501**, 217–221 (2013).
32. Vissers, L. E. L. M. et al. A de novo paradigm for mental retardation. Nat Genet **42**, 1109–1112 (2010).
33. de Ligt, J. et al. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N Engl J Med **367**, 1921–1929 (2012).
34. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet **380**, 1674–1682 (2012).
35. Chong, J. X. et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet **97**, 199–215 (2015).
36. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet **17**, 224–238 (2016).
37. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human Copy Number Variation and Complex Genetic Disease. Annu Rev Genet **45**, 203–226 (2011).
38. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of Mutation Rate Variation in the Human Germline. Annu Rev Genomics Hum Genet **15**, 47–70 (2014).
39. Korona, D. A., LeCompte, K. G. & Pursell, Z. F. The high fidelity and unique error signature of human DNA polymerase. Nucleic Acids Res **39**, 1763–1773 (2011).
40. Schmitt, M. W., Matsumoto, Y. & Loeb, L. A. High fidelity and lesion bypass capability of human DNA polymerase δ. Biochimie **91**, 1163–1172 (2009).
41. Kunkel, T. A. & Erie, D. A. Eukaryotic Mismatch Repair in Relation to DNA Replication. Annu Rev Genet **49**, 291–313 (2015).
42. Maki, H. Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses. Annu Rev Genet **36**, 279–303 (2002).
43. Lindahl, T. Quality Control by DNA Repair. Science **286**, 1897–1905 (1999).
44. Gao, Z., Wyman, Mi. J., Sella, G. & Przeworski, M. Interpreting the Dependence of Mutation Rates on Age and Time. PLOS Biol **14**, e1002355 (2016).
45. Goriely, A. & Wilkie, A. O. M. Paternal Age Effect Mutations and Selfish Spermatogonial Selection: Causes and Consequences for Human Disease. Am J Hum Genet **90**, 175–200 (2012).
46. Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. Science **349**, 1478–1483 (2015).
47. Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. Nat Genet **41**, 393–395 (2009).
48. Chen, C. L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res **20**, 447–457 (2010).
49. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet **91**, 1033–1040 (2012).
50. Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet **33**, 514–517 (2003).
51. Haradhvala, N. J. et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell **164**, 538–549 (2016).
52. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature **488**, 504–507 (2012).
53. Chan, K. & Gordenin, D. A. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. Annu Rev Genet **49**, 243–267 (2015).
54. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet **12**, 756–766 (2011).

55. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature **536**, 285–291 (2016).
56. Shendure, J. Human genomics: A deep dive into genetic variation. Nature **536**, 277–278 (2016).
57. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. Nat Rev Genet **16**, 213–223 (2015).
58. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature **500**, 415–421 (2013).
59. Petljak, M. & Alexandrov, L. B. Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis bgw055 (2016). doi:10.1093/carcin/bgw055
60. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. Nat Genet **47**, 1402–1407 (2015).
61. Sakofsky, C. J. et al. Break-induced replication is a source of mutation clusters underlying kataegis. Cell Rep **7**, 1640–1648 (2014).
62. Carvalho, C. M. B. et al. Replicative mechanisms for CNV formation are error prone. Nat Genet **45**, 1319–26 (2013).
63. Neumann, R., Lawson, V. E. & Jeffreys, A. J. Dynamics and processes of copy number instability in human -globin genes. Proc Natl Acad Sci **107**, 8304–8309 (2010).
64. Smith, D. I., Zhu, Y., McAvoy, S. & Kuhn, R. Common fragile sites, extremely large genes, neural development and cancer. Cancer Lett **232**, 48–57 (2006).
65. Stults, D. M., Killen, M. W., Pierce, H. H. & Pierce, A. J. Genomic architecture and inheritance of human ribosomal RNA gene clusters. Genome Res **18**, 13–18 (2007).
66. Bailey, J. A. Recent Segmental Duplications in the Human Genome. Science **297**, 1003–1007 (2002).
67. Weber, J. L. & Wong, C. Mutation of human short tandem repeats. Hum Mol Genet **2**, 1123–1128 (1993).
68. Sun, J. X. et al. A direct characterization of human mutation based on microsatellites. Nat Genet **44**, 1161–1165 (2012).
69. Lupski, J. R. New mutations and intellectual function. Nat Genet **42**, 1036–1038 (2010).
70. Acuna-Hidalgo, R. et al. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. Am J Hum Genet **97**, 67–74 (2015).
71. Dal, G. M. et al. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. J Med Genet **51**, 455–459 (2014).
72. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. Nat Rev Genet **14**, 307–320 (2013).
73. Campbell, I. M. et al. Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. Am J Hum Genet **95**, 173–182 (2014).
74. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet **1**, 40–47 (2000).
75. Paul, C. & Robaire, B. Ageing of the male germ line. Nat Rev Urol **10**, 227–234 (2013).
76. Goriely, A. Decoding germline de novo point mutations. Nat Genet **48**, 823–824 (2016).
77. Uchimura, A. et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res **25**, 1125–34 (2015).
78. Sherman, S. L. et al. Non-disjunction of chromosome 21 in maternal meiosis I: evidence for a maternal age-dependent mechanism involving reduced recombination. Hum Mol Genet **3**, 1529–1535 (1994).
79. Robinson, W. Maternal meiosis I non-disjunction of chromosome 15: dependence of the maternal age effect on level of recombination. Hum Mol Genet **7**, 1011–1019 (1998).
80. Wong, W. S. W. et al. New observations on maternal age effect on germline de novo mutations. Nat Commun **7**, 10486 (2016).
81. Besenbacher, S. et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun **6**, 5969 (2015).
82. Voet, T., Vanneste, E. & Vermeesch, J. R. The Human Cleavage Stage Embryo Is a Cradle of Chromosomal Rearrangements. Cytogenet Genome Res **133**, 160–168 (2011).

83. Vanneste, E. et al. Chromosome instability is common in human cleavage-stage embryos. Nat Med **15**, 577–583 (2009).
84. Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic Genome Activation During the Maternal-to-Zygotic Transition. Annu Rev Cell Dev Biol **30**, 581–613 (2014).
85. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353,** aaf7907 (2016).
86. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature **513**, 422–425 (2014).
87. Youssoufian, H. & Pyeritz, R. E. Mechanisms and consequences of somatic mosaicism in humans. Nat Rev Genet **3**, 748–758 (2002).
88. Huang, A. Y. et al. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. Cell Res **24**, 1311–1327 (2014).
89. Walter, C. A., Intano, G. W., McCarrey, J. R., McMahan, C. A. & Walter, R. B. Mutation frequency declines during spermatogenesis in young mice but increases in old mice. Proc Natl Acad Sci **95**, 10015–10019 (1998).
90. Kohler, S. W. et al. Spectra of spontaneous and mutagen-induced mutations in the lacI gene in transgenic mice. Proc Natl Acad Sci **88**, 7958–7962 (1991).
91. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci **110**, 1999–2004 (2013).
92. O'Huallachain, M., Karczewski, K. J., Weissman, S. M., Urban, A. E. & Snyder, M. P. Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci **109**, 18018–18023 (2012).
93. Laurie, C. C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet **44**, 642–650 (2012).
94. Jacobs, K. B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet **44**, 651–658 (2012).
95. Forsberg, L. A. et al. Age-Related Somatic Structural Changes in the Nuclear Genome of Human Blood Cells. Am J Hum Genet **90**, 217–228 (2012).
96. Stone, J. F. & Sandberg, A. A. Sex chromosome aneuploidy and aging. Mutat Res **338**, 107–113 (1995).
97. Dumanski, J. P. et al. Smoking is associated with mosaic loss of chromosome Y. Science **347**, 81–83 (2015).
98. Yadav, V. K., DeGregori, J. & De, S. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. Nucleic Acids Res **44**, 2075–2084 (2016).
99. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science **350**, 94–98 (2015).
100. Genovese, G. et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. N Engl J Med **371**, 2477–2487 (2014).
101. Jaiswal, S. et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N Engl J Med **371**, 2488–2498 (2014).
102. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med **20**, 1472–1478 (2014).
103. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science **348**, 880–886 (2015).
104. Dolle, M. E. T., Snyder, W. K., Gossen, J. A., Lohman, P. H. M. & Vijg, J. Distinct spectra of somatic mutations accumulated with age in mouse heart and small intestine. *Proc Natl Acad Sci* **97,** 8403–8408 (2000).
105. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science **347**, 78–81 (2015).
106. Ramsey, M. J. et al. The effects of age and lifestyle factors on the accumulation of cytogenetic damage as measured by chromosome painting. Mutat Res DNAging **338**, 95–106 (1995).
107. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529,** 43–47 (2015).
108. Sánchez Alvarado, A. & Yamanaka, S. Rethinking Differentiation: Stem Cells, Regeneration, and

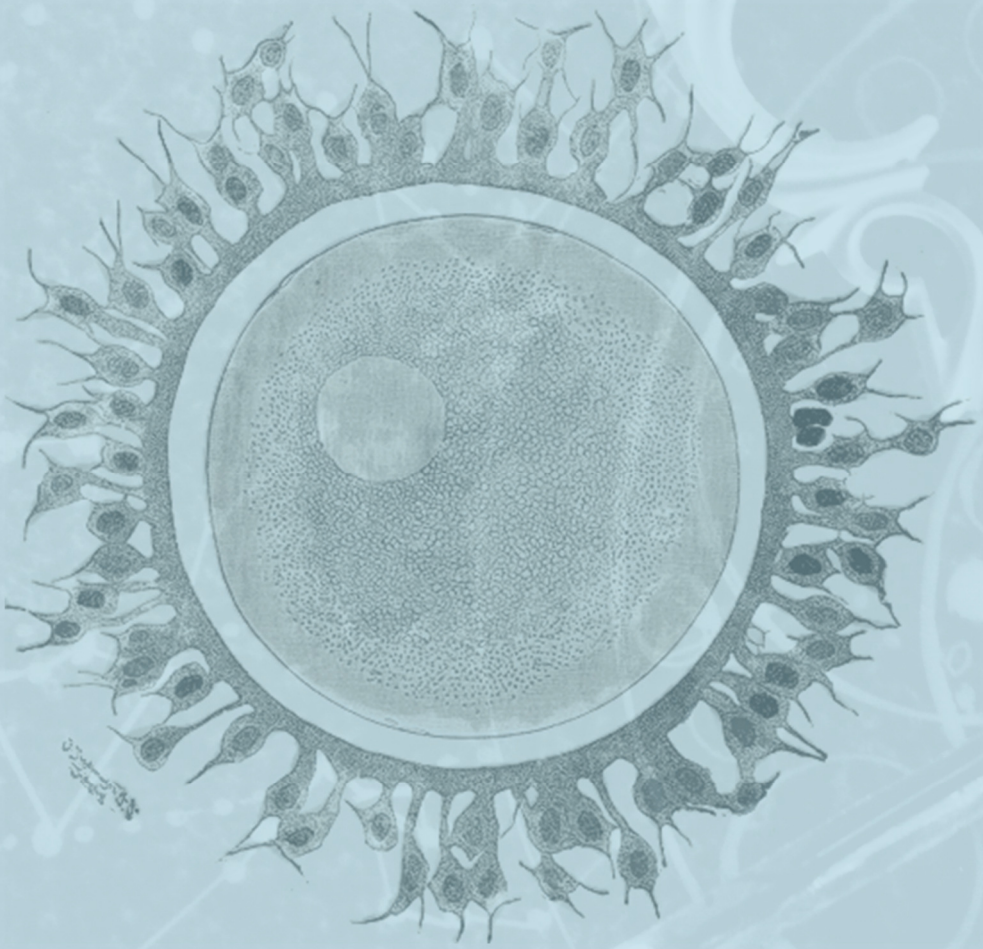
Plasticity. *Cell* **157,** 110–119 (2014).
109. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. Eur J Hum Genet **20**, 490–497 (2012).
110. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet **14**, 681–91 (2013).
111. Yang, Y. et al. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. JAMA **312**, 1870 (2014).
112. Posey, J. E. et al. Molecular diagnostic experience of whole-exome sequencing in adult patients. Genet Med **18**, 678–685 (2016).
113. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. Nat Rev Genet **17**, 9–18 (2015).
114. Baxter, a J. et al. The epidemiology and global burden of autism spectrum disorders. Psychol Med **45**, 601–613 (2015).
115. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nat Neurosci **17**, 764–772 (2014).
116. Iossifov, I. et al. Low load for disruptive mutations in autism genes and their biased transmission. Proc Natl Acad Sci **112**, E5600–E5607 (2015).
117. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. Nat Med **22**, 345–361 (2016).
118. Gratten, J. et al. Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. Nat Genet **48**, 718–724 (2016).
119. Homsy, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science **350**, 1262–1266 (2015).
120. Zaidi, S. et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature **498**, 220–223 (2013).
121. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48,** 1060–1065 (2016).
122. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet **46**, 944–950 (2014).
123. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature **519**, 223–8 (2015).
124. Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci **19**, 1194–1196 (2016).
125. Pamphlett, R., Morahan, J. M. & Yu, B. Using case-parent trios to look for rare de novo genetic variants in adult-onset neurodegenerative diseases. J Neurosci Methods **197**, 297–301 (2011).
126. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. Nature **506**, 179–184 (2014).
127. Gauthier, J. et al. De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. Proc Natl Acad Sci **107**, 7863–7868 (2010).
128. Xu, B. et al. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet **40**, 880–885 (2008).
129. Kun-Rodrigues, C. et al. A systematic screening to identify de novo mutations causing sporadic early-onset Parkinson's disease. Hum Mol Genet **24**, 6711–6720 (2015).
130. Chesi, A. et al. Exome sequencing to identify de novo mutations in sporadic ALS trios. Nat Neurosci **16**, 851–855 (2013).
131. Steinberg, K. M., Yu, B., Koboldt, D. C., Mardis, E. R. & Pamphlett, R. Exome sequencing of case-unaffected-parents trios reveals recessive and de novo genetic variants in sporadic ALS. Sci Rep **5**, 9124 (2015).
132. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. Science **349**, 1489–1494 (2015).
133. Georgieva, L. et al. De novo CNVs in bipolar affective disorder and schizophrenia. Hum Mol Genet **23**, 6677–6683 (2014).
134. Singh, T. et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci **19**, 571–7 (2016).
135. Xu, B. et al. De novo gene mutations highlight patterns of genetic and neural complexity in

schizophrenia. Nat Genet **44**, 1365–1369 (2012).
136. Zhu, X., Need, A. C., Petrovski, S. & Goldstein, D. B. One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. Nat Neurosci **17**, 773–781 (2014).
137. Jacob Gratten, Naomi R Wray, Matthew C Keller, and P. M. V. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat Neurosci **17**, 782–790. (2014).
138. van Os, J., Kenis, G. & Rutten, B. P. F. The environment and schizophrenia. Nature **468**, 203–212 (2010).
139. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic Mutation, Genomic Variation, and Neurological Disease. Science **341**, 1237758–1237758 (2013).
140. Insel, T. R. Brain somatic mutations: the dark matter of psychiatric genetics? Mol Psychiatry **19**, 156–158 (2014).
141. Gonzalez, K. D. et al. High frequency of de novo mutations in Li-Fraumeni syndrome. J Med Genet **46**, 689–693 (2009).
142. Aretz, S. et al. Frequency and parental origin of de novo APC mutations in familial adenomatous polyposis. Eur J Hum Genet **12**, 52–58 (2004).
143. Golmard, L. et al. Breast and ovarian cancer predisposition due to de novo BRCA1 and BRCA2 mutations. Oncogene **35**, 1324–1327 (2016).
144. Win, A. K. et al. Determining the frequency of de novo germline mutations in DNA mismatch repair genes. J Med Genet **48**, 530–534 (2011).
145. Hafner, C., Toll, A. & Real, F. X. HRAS Mutation Mosaicism Causing Urothelial Cancer and Epidermal Nevus. N Engl J Med **365**, 1940–1942 (2011).
146. Zhang, J. et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med **373**, 2336–2346 (2015).
147. Fernández, L. C., Torres, M. & Real, F. X. Somatic mosaicism: on the road to cancer. Nat Rev Cancer **16**, 43–55 (2015).
148. Neveling, K. et al. Next-generation genetic testing for retinitis pigmentosa. Hum Mutat **33**, 963–972 (2012).
149. Glöckle, N. et al. Panel-based next generation sequencing as a reliable and efficient technique to detect mutations in unselected patients with retinal dystrophies. Eur J Hum Genet **22**, 99–104 (2014).
150. Kim, N. K. D. et al. Whole-exome sequencing reveals diverse modes of inheritance in sporadic mild to moderate sensorineural hearing loss in a pediatric population. Genet Med **17**, 901–911 (2015).
151. Sun, C. et al. An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. Nat Genet **23**, 429–432 (1999).
152. Moro, E. Male Infertility Caused by a de Novo Partial Deletion of the DAZ Cluster on the Y Chromosome. J Clin Endocrinol Metab **85**, 4069–4073 (2000).
153. Lindhurst, M. J. et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med **365**, 611–9 (2011).
154. Shirley, M. D. et al. Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. N Engl J Med **368**, 1971–9 (2013).
155. Rivière, J.-B. et al. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat Genet **44**, 934–940 (2012).
156. Goriely, A. & Wilkie, A. O. M. Missing heritability: paternal age effect mutations and selfish spermatogonia. Nat Rev Genet **11**, 589–589 (2010).
157. Yoon, S.-R. et al. Age-Dependent Germline Mosaicism of the Most Common Noonan Syndrome Mutation Shows the Signature of Germline Selection. Am J Hum Genet **92**, 917–926 (2013).
158. Giannoulatou, E. et al. Contributions of intrinsic mutation rate and selfish selection to levels of de novo HRAS mutations in the paternal germline. Proc Natl Acad Sci **110**, 20152–20157 (2013).
159. Arnheim, N. & Calabrese, P. Germline Stem Cell Competition, Mutation Hot Spots, Genetic Disorders, and Older Fathers. Annu Rev Genomics Hum Genet **17**, (2016).
160. Maher, G. J. et al. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. Proc Natl Acad Sci **113**, 2454–2459 (2016).

161. Goriely, A. et al. Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. Nat Genet **41**, 1247–1252 (2009).
162. Goriely, A., McVean, G. A. T., Röjmyr, M., Ingemarsson, B. & Wilkie, A. O. M. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. Science **301**, 643–6 (2003).
163. Choi, S.-K., Yoon, S.-R., Calabrese, P. & Arnheim, N. Positive Selection for New Disease Mutations in the Human Germline: Evidence from the Heritable Cancer Syndrome Multiple Endocrine Neoplasia Type 2B. PLoS Genet **8**, e1002420 (2012).
164. Arnheim, N. & Calabrese, P. Understanding what determines the frequency and pattern of human germline mutations. Nat Rev Genet **10**, 478–488 (2009).
165. Goriely, A., McGrath, J. J., Hultman, C. M., Wilkie, A. O. M. & Malaspina, D. 'Selfish Spermatogonial Selection': A Novel Mechanism for the Association Between Advanced Paternal Age and Neurodevelopmental Disorders. Am J Psychiatry **170**, 599–608 (2013).
166. Huisman, S. A., Redeker, E. J. W., Maas, S. M., Mannens, M. M. & Hennekam, R. C. M. High rate of mosaicism in individuals with Cornelia de Lange syndrome. J Med Genet **50**, 339–344 (2013).
167. Halvorsen, M. et al. Mosaic mutations in early-onset genetic diseases. Genet Med **18**, 746–749 (2016).
168. Jamuar, S. S. et al. Somatic Mutations in Cerebral Cortical Malformations. N Engl J Med **371**, 733–743 (2014).
169. Okajima, K., Warman, M. L., Byrne, L. C. & Kerr, D. S. Somatic mosaicism in a male with an exon skipping mutation in PDHA1 of the pyruvate dehydrogenase complex results in a milder phenotype. Mol Genet Metab **87**, 162–8 (2006).
170. Plant, K. E., Boye, E., Green, P. M., Vetrie, D. & Flinter, F. A. Somatic mosaicism associated with a mild Alport syndrome phenotype. J Med Genet **37**, 238–9 (2000).
171. Groesser, L. et al. Postzygotic HRAS and KRAS mutations cause nevus sebaceous and Schimmelpenning syndrome. Nat Genet **44**, 783–787 (2012).
172. Happle, R. Lethal genes surviving by mosaicism: A possible explanation for sporadic birth defects involving the skin. J Am Acad Dermatol **16**, 899–906 (1987).
173. Weinstein, L. S. et al. Activating Mutations of the Stimulatory G Protein in the McCune–Albright Syndrome. N Engl J Med **325**, 1688–1695 (1991).
174. Kurek, K. C. et al. Somatic Mosaic Activating Mutations in PIK3CA Cause CLOVES Syndrome. Am J Hum Genet **90**, 1108–1115 (2012).
175. Mirzaa, G. et al. PIK3CA-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. JCI Insight **1**, 1–18 (2016).
176. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
177. Hafner, C. & Groesser, L. Mosaic RASopathies. Cell Cycle **12**, 43–50 (2013).
178. Pollock, P. M. et al. High frequency of BRAF mutations in nevi. Nat Genet **33**, 19–20 (2002).
179. Aoki, Y. et al. Germline mutations in HRAS proto-oncogene cause Costello syndrome. Nat Genet **37**, 1038–1040 (2005).
180. Levinsohn, J. L. et al. Somatic HRAS p.G12S Mutation Causes Woolly Hair and Epidermal Nevi. J Invest Dermatol **134**, 1149–1152 (2014).
181. Beukers, W., Hercegovac, A. & Zwarthoff, E. C. HRAS mutations in bladder cancer at an early age and the possible association with the Costello Syndrome. Eur J Hum Genet **22**, 837–839 (2014).
182. Luks, V. L. et al. Lymphatic and Other Vascular Malformative/Overgrowth Disorders Are Caused by Somatic Mutations in PIK3CA. J Pediatr **166**, 1048–1054.e5 (2015).
183. Limaye, N. et al. Somatic Activating PIK3CA Mutations Cause Venous Malformation. Am J Hum Genet **97**, 914–921 (2015).
184. Amoyel, M. & Bach, E. a. Cell competition: how to eliminate your neighbours. Development **141**, 988–1000 (2014).
185. Levayer, R. & Moreno, E. Mechanisms of cell competition: Themes and variations. *J Cell Biol* **200,** 689–698 (2013).
186. Clavería, C., Giovinazzo, G., Sierra, R. & Torres, M. Myc-driven endogenous cell competition in the early mammalian embryo. Nature **500**, 39–44 (2013).
187. Sancho, M. *et al.* Competitive Interactions Eliminate Unfit Embryonic Stem Cells at the Onset

of Differentiation. *Dev Cell* **26,** 19–30 (2013).
188. Wieacker, P. & Wieland, I. Clinical and genetic aspects of craniofrontonasal syndrome: Towards resolving a genetic paradox. Mol Genet Metab **86**, 110–116 (2005).
189. Dibbens, L. M. et al. X-linked protocadherin 19 mutations cause female-limited epilepsy and cognitive impairment. Nat Genet **40**, 776–781 (2008).
190. Twigg, S. R. F. et al. Cellular interference in craniofrontonasal syndrome: males mosaic for mutations in the X-linked EFNB1 gene are more severely affected than true hemizygotes. Hum Mol Genet **22**, 1654–1662 (2013).
191. Campbell, I. M. et al. Parent of origin, mosaicism, and recurrence risk: Probabilistic modeling explains the broken symmetry of transmission genetics. Am J Hum Genet **95**, 345–359 (2014).
192. Stratton, M., Campbell, P. & Futreal, A. The cancer genome. Nature **458**, 719–724 (2009).
193. Ellis, N. A., Ciocci, S. & German, J. Back mutation can produce phenotype reversion in Bloom syndrome somatic cells. Hum Genet 108**,** 167–173 (2001).
194. Jongmans, M. C. J. et al. Revertant somatic mosaicism by mitotic recombination in dyskeratosis congenita. Am J Hum Genet **90**, 426–433 (2012).
195. Hirschhorn, R. In vivo reversion to normal of inherited mutations in humans. J Med Genet **40**, 721–728 (2003).
196. Joenje, H. et al. No Title. Nat Genet **22**, 379–383 (1999).
197. Shlush, L. I. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature 506, 328–33 (2014).
198. Abdel-Wahab, O. et al. ASXL1 Mutations Promote Myeloid Transformation through Loss of PRC2-Mediated Gene Repression. Cancer Cell **22**, 180–193 (2012).
199. Abkowitz, J. L. Clone Wars — The Emergence of Neoplastic Blood-Cell Clones with Aging. N Engl J Med **371**, 2523–2525 (2014).
200. McKerrell, T. et al. Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. Cell Rep **10**, 1239–1245 (2015).
201. Beerman, I. et al. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. Proc Natl Acad Sci U S A **107**, 5465–5470 (2010).
202. Li, L. Normal Stem Cells and Cancer Stem Cells: The Niche Matters. Cancer Res **66**, 4553–4557 (2006).
203. Corces-Zimmerman, M. R. & Majeti, R. Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia* **28,** 2276–2282 (2014).
204. Vermeulen, L. et al. Defining Stem Cell Dynamics in Models of Intestinal Tumor Initiation. **267**, 263–267 (2013).
205. Klein, A. M., Nakagawa, T., Ichikawa, R., Yoshida, S. & Simons, B. D. Mouse Germ Line Stem Cells Undergo Rapid and Stochastic Turnover. *Cell Stem Cell* **7,** 214–224 (2010).
206. Ng, S. B. et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet **42**, 790–793 (2010).
207. Hoischen, A. et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet **43**, 729–731 (2011).
208. Bamshad, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet **12**, 745–755 (2011).
209. Vermeesch, J. R., Balikova, I., Schrander-Stumpel, C., Fryns, J.-P. & Devriendt, K. The causality of de novo copy number variants is overestimated. Eur J Hum Genet **19**, 1112–1113 (2011).
210. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. Nature **508**, 469–476 (2014).
211. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet **12**, 628–640 (2011).
212. Sunyaev, S. R. Inferring causality and functional significance of human coding DNA variants. Hum Mol Genet **21**, R10–R17 (2012).
213. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet **46**, 310–315 (2014).
214. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9,** e1003709

(2013).
215. Minikel, E. V. *et al.* Quantifying prion disease penetrance using large population control cohorts. *Sci Transl Med* **8,** 322ra9-322ra9 (2016).
216. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* 41111 (2016).
217. Chen, R. et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol **34**, 531–8 (2016).
218. Higurashi, N. et al. A human Dravet syndrome model from patient induced pluripotent stem cells. Mol Brain 6, 19 (2013).
219. Kuechler, A. et al. Loss-of-function variants of SETD5 cause intellectual disability and the core phenotype of microdeletion 3p25.3 syndrome. Eur J Hum Genet **23**, 753–760 (2015).
220. Lupiáñez, D. G. et al. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell **161**, 1012–1025 (2015).
221. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature **513**, 120–123 (2014).
222. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res **23**, 843–854 (2013).
223. Sun, K. et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci **112**, E5503–E5512 (2015).
224. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell **164**, 57–68 (2016).
225. Lehmann-Werman, R. et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci **113**, E1826–E1834 (2016).
226. Stessman, H. A., Bernier, R. & Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. Cell **156**, 872–877 (2014).
227. Klepper, J. et al. Seizure Control and Acceptance of the Ketogenic Diet in GLUT1 Deficiency Syndrome: A 2- to 5-Year Follow-Up of 15 Children Enrolled Prospectively. Neuropediatrics **36**, 302–308 (2005).
228. Brandler, W. M. & Sebat, J. From De Novo Mutations to Personalized Therapeutic Interventions in Autism. Annu Rev Med **66**, 487–507 (2015).
229. James, C. A., Hadley, D. W., Holtzman, N. A. & Winkelstein, J. A. How does the mode of inheritance of a genetic condition influence families? A study of guilt, blame, stigma, and understanding of inheritance and reproductive risks in families with X-linked and autosomal recessive diseases. Genet Med **8**, 234–242 (2006).
230. McAllister, M. et al. The emotional effects of genetic diseases: Implications for clinical genetics. Am J Med Genet Part A **143A**, 2651–2661 (2007).
231. Krupp, D. R. et al. Exonic somatic mutations contribute risk for autism spectrum disorder. bioRxiv (2016). doi:10.1101/083428
232. D'Onofrio, B. M. et al. Paternal Age at Childbearing and Offspring Psychiatric and Academic Morbidity. JAMA Psychiatry **71**, 432 (2014).
233. Yang, Y. et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N Engl J Med **369**, 1502–1511 (2013).
234. Cobo, A. et al. Oocyte vitrification as an efficient option for elective fertility preservation. Fertil Steril **105**, 755–764.e8 (2016).
235. Flöttmann, R. *et al.* Microdeletions on 6p22.3 are associated with mesomelic dysplasia Savarirayan type. *J Med Genet* **52,** 476–483 (2015).
236. Scally, A. Mutation rates and the evolution of germline structure. Philos Trans R Soc B Biol Sci **371**, 20150137 (2016).
237. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature **517**, 608–611 (2014).



Human oocyte. The zona pellucida is surrounded by cells of the corona radiata.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 3:
Thyroid hormone resistance syndrome due to *de novo* mutations in the thyroid hormone receptor α gene *(THRA)*

Adapted from:

Tylki-Szymańska A.*, **Acuna-Hidalgo R.***, Krajewska-Walasek M., Lecka-Ambroziak A., Steehouwer M., Gilissen C., Brunner H.G., Jurecka A., Różdżyńska-Świątkowska A., Hoischen A. & Chrzanowska K.H. Thyroid hormone resistance syndrome due to mutations in the thyroid hormone receptor α gene (*THRA*). J Med Genet 52, 312–316 (2015).

* These authors contributed equally to this work

Abstract

Resistance to thyroid hormone is characterized by a lack of response of peripheral tissues to the active form of thyroid hormone (triiodothyronine, T3). In about 85% of cases, a mutation in *THRB*, the gene coding for thyroid receptor β (TRβ), is the cause of this disorder. Recently, individual reports described the first patients with thyroid hormone receptor α gene (*THRA*) defects. We used longitudinal clinical assessments over a period of 18 years at one hospital setting combined with biochemical and molecular studies to characterize a novel thyroid hormone resistance syndrome in a cohort of six patients from five families. Using whole exome sequencing and subsequent Sanger sequencing, we identified truncating and missense mutations in the *THRA* gene in five of six individuals and describe a distinct and consistent phenotype of mild hypothyroidism (growth retardation, relatively high birth length and weight, mild to moderate intellectual disability, mild skeletal dysplasia and constipation), specific facial features (round, somewhat coarse and flat face) and macrocephaly. Laboratory investigations revealed anemia and slightly elevated cholesterol, while the thyroid profile showed low free thyroxine (fT4) levels coupled with high free T3 (fT3), leading to an altered T4:T3 ratio, along with normal thyroid-stimulating hormone levels. We observed a genotype-phenotype correlation, with milder outcomes for missense mutations and more severe phenotypical effects for truncating mutations. In patients with clinical symptoms of mild hypothyroidism without confirmation in endocrine studies, a molecular study of *THRA* defects is strongly recommended.

Introduction

Thyroid hormones (THs, comprising triiodothyronine – T3 and thyroxine – T4) have diverse actions, which include the regulation of skeletal growth, maturation of the central nervous system, cardiac and gastrointestinal function and energy homeostasis.[1] Thyroid hormones exert their effects through alpha (TRα) and beta (TRβ) receptors,[1] which belong to the nuclear receptors superfamily including receptors for gluco- and mineralocorticoids, estrogens, progesterone and vitamin D. TRα and TRβ are encoded by two genes (*THRA* and *THRB*), each of which undergoes alternate splicing to generate receptor subtypes (TRα1, TRα2, TRβ1, TRβ2, and TRβ3) with distinct tissue distribution.[1] TRα1 and TRα2 are located mainly in bones, digestive tract, cardiac and skeletal muscle and central nervous system.[2] Three subtypes of TRβ are mainly expressed in the liver, kidney, hypothalamus and pituitary, where they regulate the release of thyrotropin, subject to the current concentration of THs in blood and in the thyroid gland.[3,4]

Resistance to thyroid hormone (RTH) is characterized by a lack of response of peripheral tissues to the active form of thyroid hormone (T3). Two forms of inheritable RTH have been described, leading to clinically distinct phenotypes: RTH β and α. In more than 85% of the cases, RTH β is caused by a mutation in *THRB*, the gene coding for TRβ. This disorder affects 1 per 50,000 live births and can appear both sporadically, as well as in an inherited autosomal recessive or dominant manner.[5] Affected individuals clinically present with goiter, short stature, decreased body mass, enlarged heart and mild psychomotor retardation. Furthermore, the disruption of TRβ function in the hypothalamus, pituitary and thyroid gland results in a loss of the negative feedback in this axis. This is reflected in the thyroid function tests in affected individuals, which show increased levels of T4 and T3 coupled to inappropriately normal levels of thyroid-stimulating hormone (TSH).[6] Approximately 15% of patients with RTH do not harbor a mutation in the *THRB* gene, which has led to the suspicion that this disorder may be caused by alterations in other thyroid hormone receptor-related genes. These candidate genes include co-regulators, co-activators and co-repressors

regulating TRβ function. However, efforts to screen individuals with RTH for mutations in these genes have failed so far.[7]

Particularly, the absence of mutations in *THRA* in patients with RTH has been met with puzzlement, leading to speculation that mutations in this gene could be extremely rare or undiagnosed. Recently, three case reports described the first patients with heterozygous truncating mutations in the last exon of *THRA* presenting clinical features suggestive of thyroid hormone resistance, however different from RTH β.[1,5,8,9]

Here, we present the clinical and laboratory features of five pediatric patients (one male and four females, age 4 to 18 years) as well as one adult patient (age 39 years) with RTH α, a novel thyroid hormone resistance syndrome due to truncating and missense *THRA* mutations. To our knowledge, this is the first study with a larger number of patients and a description of a distinct and consistent phenotype of mild hypothyroidism, associated with macrocephaly, specific facial features, skeletal abnormalities and altered T4:T3 ratio, along with normal TSH levels.

Results

Clinical syndrome

The clinical syndrome was similar in all patients and occurred due to *de novo* mutations or was transmitted as a dominant trait in one family (Supplementary Figure S1). Supplementary Table S1 summarizes the clinical data for the six patients studied. All patients presented with a similar clinical outcome and phenotype which comprises growth retardation (with relatively short limbs, hands and feet, and long thorax), mild to moderate intellectual disability, mild skeletal dysplasia, constipation, deep voice and specific facial features (round/puffy, somewhat coarse, flat face, flat nasal bridge, upturned nose, hypertelorism,) with relative macrocephaly (Figure 1 and 2). Other changes included puffy hands and feet, club feet, tortuosity of arteries of the dorsal area of the feet and hands, and skin with rough and doughy texture (Figure 3). In two patients with more pronounced clinical features mild cardiomyopathy was also noted. From infancy, all patients presented mild to moderate motor and mental retardation, deep voice and constipation.

The developmental phenotype was characterized by delayed delivery after week 40, relatively high birth weight and length, large head circumference relation to the chest circumference and floppiness (Supplementary Table S1). Newborn screening allowed exclusion of hypothyroidism. From the age of 2 years, the height gain became slower, leading to the marked short stature at the later age. Growth retardation was disproportionate, affecting the lower segment

more than the upper segment. The Polish reference charts for sitting height are available for children aged >4 years. The following indexes were calculated:

1. Trunk-lower extremities index (according to Giuffrida-Ruggeri) using the following formula:

$$\frac{sitting\ height}{body\ height} \times 100$$

2. Skelic index (according to Manouvrier) using following formula:

$$\frac{body\ height - sitting\ height}{sitting\ height} \times 100$$

Trunk-lower extremities index in all patients except one (patient 1 after puberty had skeletal deformity) revealed long trunk. Skelic index revealed all patients to be hyperbrachyskelic.

Figure 1. Progression of facial features in five patients with a novel thyroid hormone resistance syndrome due to truncating mutations in *THRA*. Patient 1, aged 2, 2, 7 and 18 years, respectively (1), patient 2, aged 2.5 and 15 years, respectively (2); patient 3, aged 18 months, 4 and 6 years, respectively (3); father of patient 3, aged 39 years (F); patient 4, aged 2 and 5 years (4); patient 5, aged 3 months, 2, 4 and 5 years, respectively (5). Note the characteristic facial traits such as macrocephaly, round/puffy, somewhat coarse and flat face, flat nasal bridge, upturned nose and hypertelorism.

Figure 2. Photo gallery of five patients with a novel thyroid hormone resistance syndrome due to truncating mutations in *THRA*. Patient 1, aged 2 (1) and 18 (1a) years; patient 2, aged 2.5 years (2); patient 3, aged 6 years (3); father of patient, aged 39 years (F); patient 4, aged 5 years (4) and patient 5, aged 5 years (5). Note the short stature, relatively short limbs, hands and feet and long thorax.

Figure 3. Hands and feet in a novel thyroid hormone resistance syndrome due to truncating mutations in *THRA*. Please note puffy hands and feet, club feet, tortuosity of arteries of the dorsal area of the feet and hands and skin with rough and doughy texture.

Laboratory investigations revealed anemia (low red blood cells and hemoglobin level), slightly elevated creatine kinase and cholesterol. The thyroid profile showed low free thyroxine (fT4) levels coupled with high free T3 (fT3), leading to an altered T4:T3 ratio, along with normal TSH levels (Supplementary Table S1 and S2).

X-ray examination revealed ovoid immature form of the vertebral bodies, anterior superior ossification defect in the lower thoracic and upper lumbar bodies, and hypoplasia of the acetabular and supraacetabular portions of the ilia and *coxa vara*. X-rays of hands showed minimal changes in the tubular bones of the hands, which became abnormally short, wide and deformed. A consulting radiologist suggested similarities to the radiological characteristics of hypothyroidism (Supplementary Figure S2).

Management

Attempts to treat affected patients with levothyroxine at standard doses did not improve any of the clinical symptoms.

Genetic studies

The two patients with the most severe phenotype were selected for exome sequencing (patient 1 and 2, Supplementary Table S1). Both affected individuals were found to have heterozygous stop mutations in the last exon of the coding sequence for *THRA* isoform 1 (RefSeq NM_199334; c.1176C>A, [p.C392X] in patient 1 and c.1207G>T, [p.E403X] in patient 2. These mutations were absent in in-house exomes (2,094 exomes), as well as 6,500 public exomes from Exome Variant Server (EVS, http://evs.gs.washington.edu/EVS). Truncating mutations in the last exon of *THRA* isoform 1 were previously reported to cause a similar phenotype in three individuals from two families.[1] We validated these mutations by Sanger sequencing and confirmed that they had occurred *de novo* in the patients. Using Sanger sequencing, we screened the rest of our cohort for mutations in the coding regions of *THRA.* We identified a heterozygous missense mutation in the last exon of *THRA* isoform 1 in two additional unrelated individuals (c.1207G>A, [p.E403K] and c.1193C>G, [p.P398R] in patient 3 and 4, respectively). The mutation in patient 3 was inherited from the father, who was also affected, while the mutation in patient 4 occurred *de novo*. All of these mutations cluster to the C-terminus of the TRα1 protein, leaving TRα2 unaltered (Supplementary Figure S3), which is consistent with the previous case reports.[1,8]

Despite sharing the same clinical features, no mutations in *THRA* were found in patient 5 after screening by Sanger sequencing. We performed additional genetic studies to identify the disease-causing genetic lesion in this individual, including exome sequencing and a high-resolution microarray CNV

analysis. The absence of mutations in the *THRA* gene was confirmed by exome sequencing and no potentially pathogenic variants were identified in any candidate genes (including genes coding for TR co-factors and other proteins involved in thyroid hormone metabolism and function). Furthermore, a high-resolution microarray failed to show potentially pathogenic copy number variations. Therefore, despite genome-wide screening by different methods, we failed to identify the disease-causing genetic hit in patient 5.

Discussion

We describe a novel clinical and biochemical phenotype associated with truncating mutations in the *THRA* gene. While THRS due to mutations in *THRB* gene is a well known disorder, patients with THRS caused by mutations in *THRA* had not, until recently, been identified. As no patients with classic THRS have been found to have mutations in *THRA*, it has been often suggested that mutations in *THRA* may be extremely rare or clinically unrecognized. However, our study suggests that milder clinical features and thyroid profile in patients with this novel form of RTH α may be the major cause. This discrepancy is most likely due to the differences in tissue expression of *THRA* and *THRB*; *THRA* is expressed in the cardiac and skeletal muscle, digestive tract, bones and brain, while *THRB* is found in the liver, kidneys, hypothalamus, hypophysis and thyroid gland. In both forms of RTH, organs expressing a dysfunctional thyroid hormone receptor will respond in a tissue-specific way to the lack of thyroid hormone stimulus. Mutations in *THRB* lead to selective pituitary resistance to thyroid hormone, releasing the negative feedback in the hypothalamic-pituitary-thyroid axis, which finally results in an increase in production of thyroid hormone by the thyroid gland. The clinical picture of individuals with RTH β often includes a mixture of features of hyperthyroidism, such as tachycardia or weight loss, and of classic hypothyroidism (for instance, delayed growth or alterations in bone ossification). Interestingly, in RTH α, although the uptake of thyroid hormones is altered in several target organs, the signalling in the hypothalamus-pituitary-thyroid axis is preserved because of the regular function of TRβ2 in the hypothalamus and pituitary gland. As a result, despite the deprivation of thyroid hormone stimulus on peripheral tissues, the concentration of TSH in these individuals is found within the normal range. Despite normal/high T_3 blood serum concentration, the patients present symptoms of hypothyroidism as the result of an impaired TRα function in T_3 target tissues. All patients described so far with RTH α presented with clinical features resembling those of untreated mild congenital hypothyroidism such as mild to moderate intellectual disability, short stature, alterations in bone ossification, macroglossia and chronic constipation.[1,8,11] Our study revealed additional hallmark features of RTH α such as macrocephaly, specific facial features, skeletal abnormalities (planovalgus foot, sandal gap deformities, club foot or spine deformities), mild features of skeletal dysplasia characteristic for hypothyroidism, anaemia, elevated creatine kinase and

moderate hypercholesterolemia. Additionally, our findings suggest a certain degree of correlation between the genotype and phenotype, as patients with nonsense mutations have intellectual disability, while patients with missense mutations show low IQ levels that are still within normal ranges. In line with this, the only inherited disorder was due to a *THRA* missense mutation, while all other mutations leading to more severe outcome were due to *de novo* mutations.

Although the clinical picture of RTH α reminds of hypothyroidism, the laboratory data in these patients is inconsistent. The results of laboratory analyses in the study group are not typical for hypothyroidism. Despite a low fT_4 level, the concentration of fT_3, physiologically the strongest of the thyroid hormones, is high. Such correlation between the free thyroid hormones results from a TRα mutation, which prevents complete utilisation of thyroid hormones at the cellular level. Free thyroid hormones fall within a broad, normal range. This results in a thyroid laboratory profile, which without detailed analysis, is unremarkable for thyroid disease and thus, hampers the correct diagnosis in these patients. The above may lead to a discontinuation of attempts to find causes of hypothyroidism and a misdiagnosis. However, if results of thyroid hormone investigations show normal levels of TSH with low-normal T4 and high-normal T3 resulting in a low T4/T3 ratio, the possibility of deficit of TRα should be taken into consideration. Newborn screening for hypothyroidism, which is run in many countries does not reveal TRα deficit. We strongly recommend that patients with some clinical signs/symptoms of hypothyroidism and without confirmation in routine endocrine studies should have a molecular study of *THRA* defects.

The clustering of all the mutations identified to the last exon of *THRA* supports that the mechanism through which this mutation leads to pathogenicity is by acting as a dominant-negative TRα1 protein constitutively binding a co-repressor.[1] Remarkably, one of the missense mutations identified in our cohort, TRα1 P398R, corresponds to a mutation of the homologous residue in TRβ1 (P452R), which has been previously identified to lead to RTHβ.[12] The presence of equivalent pathogenic mutations in both TH receptors points to a common molecular mechanism underlying RTH in RTHα and RTHβ. Regarding the stop mutations in *THRA*, their localisation to the last exon allows them to avoid mRNA nonsense-mediated decay, permitting transcription of the protein. In our cohort, the two patients with nonsense mutations in *THRA* had the most severe clinical phenotypes accompanied by the largest deviations from the norm in the results of thyroid-related laboratory tests. At the same time, these two patients are the oldest in the observed group and their clinical symptoms seem to be intensifying with age. However, one can speculate that, despite the wide spectrum of clinical expression in dominant inheritance, we observe a genotype/phenotype correlation in our cohort.

Despite presenting with the same clinical phenotype, the absence of mutations in *THRA* in one patient in our cohort suggests genetic heterogeneity in this syndrome. Additional genome-wide studies failed to identify the genetic

lesion responsible for the phenotype in this patient. This is reminiscent of RTH β in which 85% of affected individuals have mutations in *THRB,* whilst mutations in this gene have been excluded in the remaining 15% without thereby identifying the disease-causing genetic lesion.[7] It has been hypothesized that individuals with so-called non-TR-RTH may have a mutation in a gene involved in TR function or TH metabolism. We speculate that this may also be the case for RTH α; a mutation or small deletion in a gene involved in TR function or TH metabolism may be the cause of the disease in patient 5. Our results suggest that this novel syndrome of thyroid hormone resistance is genetically heterogeneous and there may be additional genes involved in the pathogenesis of this disorder.

The results from studies performed on mutant *Thra* mouse models suggest that the recruitment or action of co-regulators, mainly co-repressors, on TRα1 could be a target for treatment in individuals with *THRA* mutations.[12] A better understanding of the role that TRα1 co-regulator dysfunction plays in this genetic syndrome is necessary. However, pharmaceutical compounds targeting such nuclear receptor co-regulators could be a promising treatment, not only for rare genetic syndromes such as RTH, but also for hormone-dependent cancers.[13]

Conclusions

1. Truncating mutations in the *THRA* gene lead to RTH α, a distinct and consistent phenotype of mild hypothyroidism associated with macrocephaly, specific facial features and skeletal abnormalities.
2. *THRA* mutations may be more common than expected and may escape standard laboratory/hormone level tests.
3. In patients with clinical signs/symptoms of mild hypothyroidism, without confirmation in endocrine studies, a molecular study of defects in *THRA* is recommended.

Materials and methods

Patients

All patients (n=6, age range 2 to 39 years, mean age 16, median age 11.5) were admitted and followed at The Children's Memorial Health Institute (Warsaw, Poland). All five pediatric patients were selected based on strikingly similar dysmorphic features along with developmental and metabolic alterations, suggestive for a common underlying genetic defect.

Molecular genetic studies

Blood and/or saliva samples from the patients and their parents were obtained, from which DNA was extracted and purified. Exome libraries were prepared using a SureSelect human exome v2 kit (Agilent, Santa Clara, USA) and sequencing was performed on patient DNA on a SOLiD4 platform (Life Technologies, Foster City, USA). Reads were mapped to the hg19 reference genome and variant calling was done with Lifescope software v2.1 (Life Technologies, Foster City, USA), Variant annotation, filtering and prioritization was performed as described previously.[10] Sanger sequencing was performed to validate candidate variants and prove *de novo* occurrence or show segregation.

Ethical considerations

The protocol was approved by the human-subjects institutional review board at the Children's Memorial Health Institute. The study was designed and conducted in compliance with the principles of the International Conference on Harmonisation of Technical Requirements for registration of Pharmaceuticals for Human Use Guidelines for Good Clinical Practice. Written informed consents for the genetic investigations and picture publication were provided by the parents or legal guardians.

Web resources

Exome Variant Server: http://evs.gs.washington.edu/EVS/

References

1. Bochukova E. et al. A mutation in the thyroid hormone receptor alpha gene. N Engl J Med **366**, 243–249 (2012).
2. Aranda A., Alonso-Merino E., Zambrano A. Receptors of thyroid hormones. Pediatr Endocrinol Rev **11**, 2–13 (2013)
3. Ocasio CA & Scanlan TS. Design and characterization of a thyroid hormone receptor alpha (TRalpha)-specific agonist. ACS Chem Biol **1**, 585–593 (2006).
4. Olateju T.O. & Vanderpump M.P. Thyroid hormone resistance. Ann Clin Biochem **43**, Pt 6, 431–440 (2006).
5. Reutrakul S et al. Search for abnormalities of nuclear corepressors, coactivators, and a coregulator in families with resistance to thyroid hormone without mutations in thyroid hormone receptor beta or alpha genes. J Clin Endocrinol Metab **8**,3609–3617 (2000).
6. Lazar M.A. Thyroid hormone receptors: multiple forms, multiple possibilities. Endocr Rev **14**, 184–193 (1993).
7. Refetoff S. & Dumitrescu A.M. Syndromes of reduced sensitivity to thyroid hormone: genetic defects in hormone receptors, cell transporters and deiodination. Best Pract Res Clin Endocrinol Metab **21**, 277–305 (2007).
8. van Mullem A. et al. Clinical phenotype and mutant TRalpha1. N Engl JMed **366**, 1451–1453 (2012).
9. Moran C. et al. Resistance to thyroid hormone caused by a mutation in thyroid hormone receptor (TR)alpha1 and TRalpha2: clinical, biochemical, and genetic analyses of three related patients. Lancet Diabetes Endocrinol **2**, 619–626 (2014).
10. Hoischen A. et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet **43**, 729–731 (2011).
11. Moran C. et al. An adult female with resistance to thyroid hormone mediated by defective thyroid hormone receptor alpha. J Clin Endocrinol Metab **98**, 4254–4261 (2013).
12. Amor A.J. et al. Identification of four novel mutations in the thyroid hormone receptor-beta gene in 164 Spanish and 2 Greek patients with resistance to thyroid hormone. Hormones **13**, 74–78 (2014).
13. Fozzatti L. et al. Nuclear receptor corepressor (NCOR1) regulates in vivo actions of a mutated thyroid hormone receptor alpha. Proc Natl Acad Sci USA 110,7850–7855 (2013).
14. Lonard D.M. & O'Malley B.W. Nuclear receptor coregulators: modulators of pathology and therapeutic targets. Nat Rev Endocrinol **8**, 598–604 (2012).

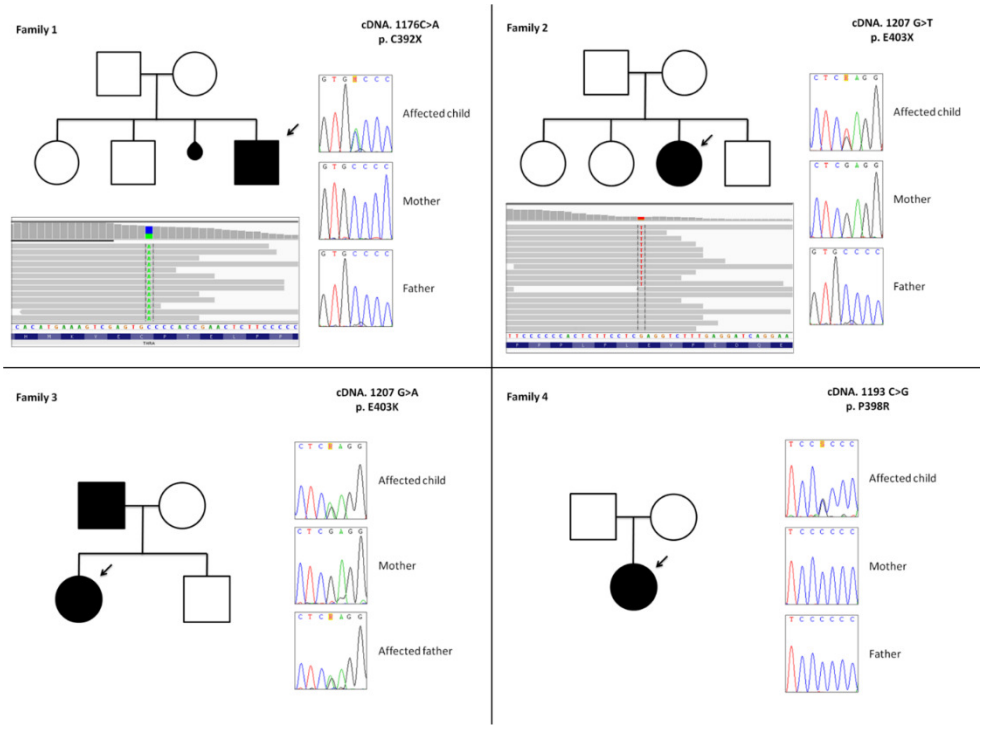
Supplementary data

Supplementary materials

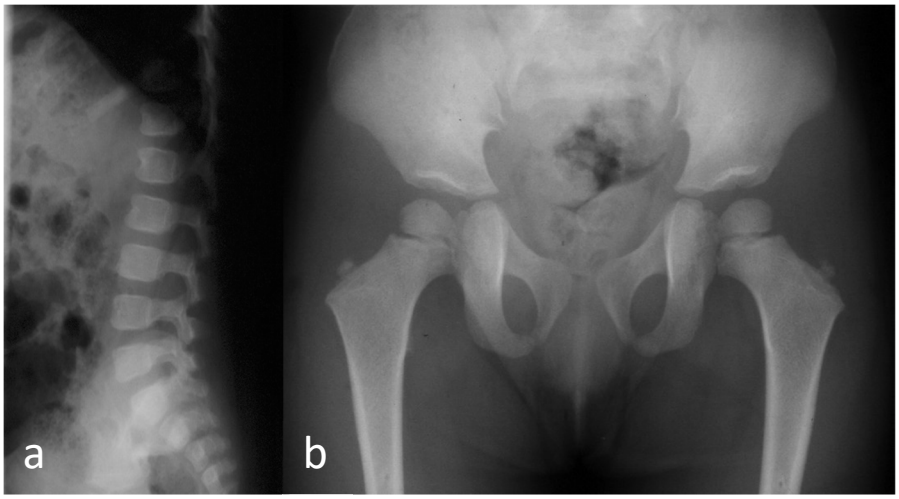
Based on the similarity and severity of their clinical features, two affected individuals were selected for exome sequencing. Blood samples were taken and DNA was extracted and purified using the QIAamp genomic DNA kit (Qiagen, Hilden, Germany). Exome libraries of patients 1 and 2 were prepared using the SureSelect v.2, 50Mb kit (Agilent) and sequenced on a SOLiD™4 platform (Life Technologies, Foster City, USA. The sequence reads were aligned to the hg19 reference genome resulting in 2.9 Gb of mapped sequences in both cases which correlates to 58-fold and 59-fold average coverage respectively (see Supplementary Table S3). Variants were called and annotated using Lifescope v.2.1. software. Called variants were prioritized based on the quality of the sequencing of each called base pair, on the predicted consequences at the protein level and taking into consideration the overlap with common variation (in-house variant database consisting of 2,094 exomes, dbSNP v.132, 1000 Genomes project and the exome variant server (EVS); Supplementary Table S4). After these prioritization steps, we performed an overlap analysis to determine in which genes were private variants identified in both of the sequenced individuals (Supplementary Table S5). In the exome data, we identified in each individual a truncating variant in *THRA* in each individual (Supplementary Figure S1 and Supplementary Table S5) and a private missense mutation in *KRT2* (which was shown to be false-positive by Sanger sequencing). We performed Sanger sequencing of the *THRA* gene from genomic DNA of the affected individuals and their parents, to check for segregation of the mutations (Supplementary Figure S1). All primers were obtained from IDT and their sequences are provided in Supplementary Table S6.

The exome of patient 5 was prepared using SureSelect v.4 (Agilent) and sequenced on a SOLID™ 5500XL Platform (Life Technologies, Foster City, USA). Mapping and variant calling was done as mentioned previously. Called variants were filtered based on the quality of the sequencing (more than 3 variant reads and variant present in more than 20% of the reads), the frequency of the variant in sequencing databases (minor allele frequency less than 0.1% and found in less than 5 samples from our in-house database) and the predicted consequences at the protein level (conserving missense and truncating coding mutations and mutations affecting the canonical splice sites). No mutations were seen in the *THRA* gene, in genes coding for known co-factors, co-regulators or co-repressors of the THRA protein nor in the genes coding for deiodinases.

Supplementary figures



Supplementary Figure S1. Pedigrees of the different families, showing sequencing data from exome sequencing and capillary electrophoresis. A c.1176C>A mutation in *THRA* was identified by exome sequencing in the index case from family 1 (marked with an arrow). This truncating mutation in THRA [p.C392X] was found to be *de novo*. Mutation c.1207G>T in the index case of family was also identified by exome sequencing. Capillary electrophoresis shows that this truncating mutation occurred *de novo* in the affected individual. The mutation in the index case of family 3, *THRA* c.1207G>A [p.E403K] was inherited from the father, who was also affected. This mutation was identified by screening through capillary electrophoresis. Screening for mutations in *THRA* by capillary electrophoresis also revealed a missense mutation (*THRA* c.1193C>G [p.P398R]) in the index case of family 4. The mutation in this patient occurred *de novo*.



Supplementary Figure S2: Radiologic finding in a novel thyroid hormone resistance syndrome due to truncating mutations in *THRA* gene. a: ovoid (immature) form of the vertebral bodies, anterosuperior ossification defect in the lower thoracic and upper lumbar bodies. b: Hypoplasia of the acetabular and supraacetabular portions of the ilia; *coxa valga*.

Supplementary Figure S3. Diagram of *THRA* isoform 1 mRNA (a) and protein (b) showing the mutations at the cDNA and protein level, which are only found in *THRA* isoform 1 and not in *THRA* isoform 2 (c and d, for mRNA and protein level respectively). Truncating mutations are shown in red triangles and missense mutations are shown in pink triangles. Dark blue rectangles represent exons that are found both in *THRA* isoform 1 and 2, while light blue rectangles represent exons that are differentially spliced in each isoform. Important domains in the THRA protein are shown in color with green representing the nuclear receptor domain and purple represents the ligand-binding domain.

Supplementary tables

Patients					
	1	**2**	**3**	**4**	**5**
Mutation (cDNA)	*THRA* C1176A	*THRA* G1207T	*THRA* G1207A	*THRA* C1193G	Not identified
Substitution (protein)	THRA C392X	THRA E403X	THRA E403K	THRA P398R	Not identified
Inheritance	*de novo*	*de novo*	inherited	*de novo*	-
Age	18	14	12	8	5
Dysmorphic features					
Macrocephaly	**+**	**+**	**+**	-	**+**
Coarse face	+	-	-	-	-
Wide forehead	+	+	-	+	-
Hypertelorism	**+**	**+**	**+**	**+**	**+**
Palpebral ptosis	**+**	**+**	**+**	-	**+**
Low or flat nasal bridge	**+**	**+**	**+**	**+**	**+**
Micrognatia	+	-	+	-	-
Macroglossia	-	+	-	-	-
Short neck	+	-	+	-	+
Growth and skeletal abnormalities					
Delayed growth	+	+	-	+	-
Short stature	**+**	**+**	**+**	-	**+**
Elongated thorax	**+**	**+**	**+**	**+**	-
Lumbar kyphosis	+	+	-	-	-
Short limbs	**+**	**+**	**+**	**+**	**+**
Developmental abnormalities					
Intellectual disability	Severe	Moderate	Mild	Normal	Normal
IQ	22	70	80	95	110
Delayed speech development	+	-	-	+	-
Psychomotor alterations	**+**	**+**	**+**	**+**	-
Gestational age at birth	**41 weeks**	**42 weeks**	**42 weeks**	**42 weeks**	**40 weeks**
Birth weight	4200g (90th perc)	3950g (90th perc)	3650g (>75th perc)	3450g (>50th perc)	3920g (>75th perc)
Birth length	56cm (>98th perc)	50cm (>50th perc)	55cm (>98th perc)	55cm (>98th perc)	51 cm (>75th perc)
Head circumference	**37cm (>98th perc)**	**39cm (>98th perc)**	**39 cm (>98th perc)**	35 cm (>75th perc)	**36 cm (>98th perc)**

	Patients				
	1	**2**	**3**	**4**	**5**
Mutation (cDNA)	*THRA* C1176A	*THRA* G1207T	*THRA* G1207A	*THRA* C1193G	Not identified
Substitution (protein)	THRA C392X	THRA E403X	THRA E403K	THRA P398R	Not identified
Inheritance	*de novo*	*de novo*	inherited	*de novo*	-
Age	18	14	12	8	5
Other characteristics					
Deep or hoarse voice	**+**	**+**	**+**	+	**+**
Pale and doughy skin	**+**	**+**	**-**	-	**-**
Constipation	**+**	**+**	**+**	+	**+**
Anemia	+	+	+	+	**+**
High cholesterol	**+**	**+**	**+**	+	**-**
Thyroid laboratory tests					
fT4	**Low**	**Low/Normal**	**Low/Normal**	**Low/Normal**	Normal
fT3	**High/Upper limit**	**Upper limit**	**Upper limit**	**Upper limit**	**Upper limit**
TSH	**Normal**	**Normal**	**Normal**	Low/ Normal	**Normal**
Other thyroid function features	Negative anti-TPO and anti-TG antibodies	N/A	Limit levels for anti-TPO and anti-TG antibodies	N/A	Normal levels of thyroxine binding globulin
X-ray abnormalities	+	+	+	N/A	N/A

Supplementary Table S1: Clinical phenotype of patients with a syndrome of thyroid hormone resistance due to mutations in the *THRA* gene. + present; - not present; N/A not available. All features that are present in four or more unrelated subjects of our cohort are highlighted in bold font. The full version of this table is available at http://jmg.bmj.com/content/52/5/312.long

	THRS due to *THRA* mutations	THRS due to *THRB* mutations
fT4	Low or normal	High
fT3	Upper limit or high	High
T3:T4 ratio	Altered	High
TSH	Low or normal	High
Goiter	Not present	+
Other	Negative thyroperoxidase or thyroglobulin antibodies	Negative thyroperoxidase or thyroglobulin antibodies

Supplementary Table S2. Summary of clinical and laboratory findings relative to thyroid function in individuals with thyroid hormone resistance syndrome (THRS) due to mutations in *THRA* compared to those of individuals with THRS due to mutations in *THRB*.

	Patient 1	Patient 2	Patient 5
Total mapped bases [hg19]	2,923,298,920	2,976,245,643.00	3,628,016,187
Median target coverage	55.4	59.4	65.3
Average target coverage	58.2	58	65.4

Supplementary Table S3. Exome mapping statistics for patients on which exome sequencing was performed. The exomes from patient 1 and 2 were enriched with the SureSelect human exome v2 kit and sequenced using a SOLiD4 platform, while the exome from patient 5 was based on v4 exome data and SOLiD 5500XL sequencing.

		Patient 1	Patient 2	Patient 5
All variants		32,904	35,635	48,207
of which	Coding and splice site variants	13,894	14,685	17,150
	Private variants (i.e. not present in dbSNP132 nor in 2094 in-house exomes)	329	368	182
	Non-synonymous	252	274	125
	Variants found in genes overlapping in affected individuals	2		-

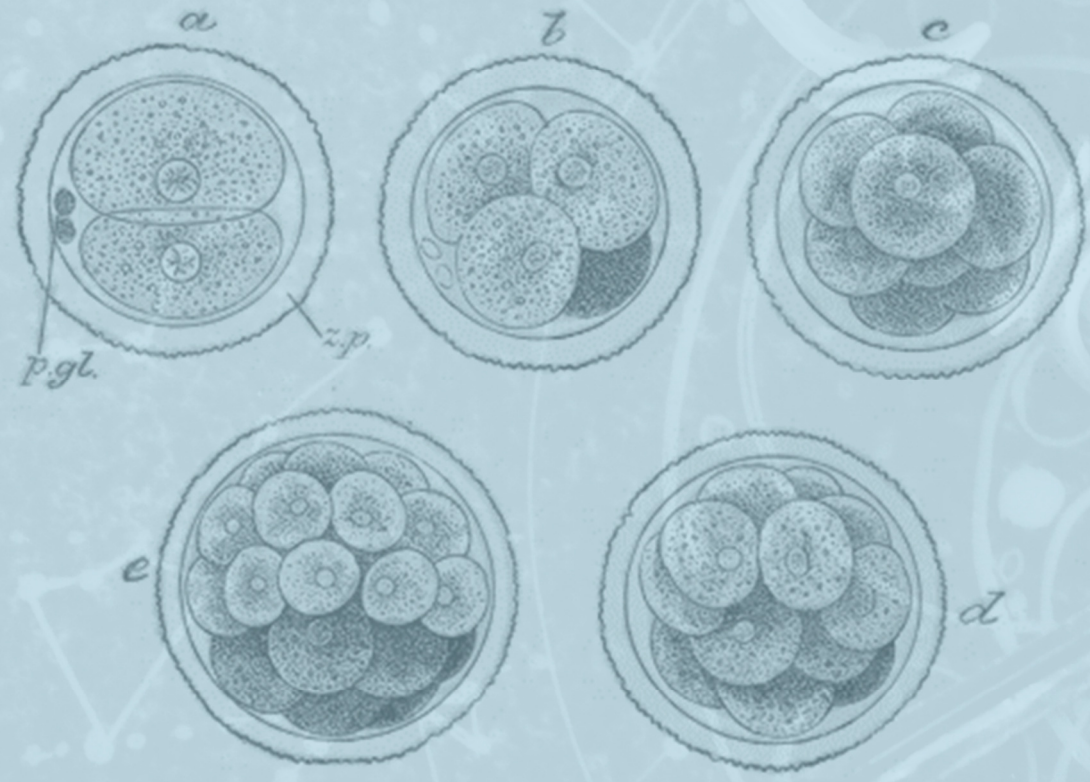
Supplementary Table S4. Variant filtering of exome data of patients 1, 2 and 5 (prioritization analysis, similar to Hoischen *et al.*[8])

Patient	genomic position [hg19]	ref	mut	% variation	Gene name	Amino acid changes	cDNA changes	phyloP	Grantham Score	Confirmed	*de novo*
1	chr12:53044330	A	C	33.3	KRT2	F198C	593A>C	1.751	205	No	-
1	chr17:38245683	G	T	63.2	THRA	E403*	1207G>T	5.444	-	Yes	Yes
2	chr12:53039095	C	T	20	KRT2	G543D	1628C>T	2.377	94	No	-
2	chr17:38245652	C	A	44	THRA	C392*	1176C>A	1.642	-	Yes	Yes

Supplementary Table S5: Private variants affecting the same gene in patient 1 and patient 2. By using the previously described variant prioritization method, we determined genes in which both affected individuals presented non-synonymous variants that had not been identified previously in controls (private variants). The sequenced individuals presented private variants in two genes, *KRT2* and *THRA*. Only the mutations in *THRA* were confirmed by Sanger sequencing.

Primer name	5'-Sequence-3'
THRA Exon 2 F	CCCTTTTAAGCATTGTCAGC
THRA Exon 2 R	CTGGGCACATCCCACATTAC
THRA Exon 3 F	TTAGGGCAGAGATGGACAGG
THRA Exon 3 R	ATCACTTGAACCTGAAGGCG
THRA Exon 4 F	CCCTCACTGATCTTGCCTTC
THRA Exon 4 R	AGGACAGCAAGGACCAAGAC
THRA Exon 5 F	CCTCGGTTTCTCCAACCTG
THRA Exon 5 R	TGGACAGCAAAAGTGTGAAGAG
THRA Exon 6 F	CTTTGGGCCTGGGACTC
THRA Exon 6 R	TGCTGGGTGTATGTGTATGC
THRA Exon 7 F	CGTTCAGAGTCCATGGGG
THRA Exon 7 R	AGCCCAGAAGAAGATTCAGG
THRA Exon 8 F	TCCTGTCCACGTCTCTCAGG
THRA Exon 8 R	ATCCAAGGACGAGAGGAAGG
THRA Exon 9 F	GGGGCCAGAGGCTCATC
THRA Exon 9 R	CACCTCCTGCTCTTGGGG

Supplementary Table S6. Sequences of the primers used to screen for mutations in *THRA* in our cohort.



First stages of division of mammalian embryos. a. Two cell stage; b. Four cell stage; c. Eight cell stage; d,e. Morula stage. z.p. zona pellucida; p.gl. polar bodies.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 4:
Postzygotic point mutations are an underrecognized source of novel genomic variation

Published as:

Acuna-Hidalgo R., Bo T., Kwint M.P., van de Vorst M., Pinelli M., Veltman J.A., Hoischen A., Vissers L.E.L.M. & Gilissen C. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. Am J Hum Genet 97, 67–74 (2015).

Abstract

De novo mutations are recognized both as an important source of genetic variation and as a prominent cause of sporadic disease in humans. Mutations identified as *de novo* are generally assumed to have occurred during gametogenesis and, consequently, to be present as germline events in an individual. As Sanger sequencing does not provide the sensitivity to reliably distinguish somatic from germline mutations, the proportion of *de novo* mutations that occurs somatically rather than in the germline remains largely unknown. To determine the contribution of postzygotic events to *de novo* mutations, we analyzed a set of 107 *de novo* mutations in 50 parent-offspring trios. Using four different sequencing techniques, we found that 7 (6.5%) of these presumed germline *de novo* mutations were in fact present as mosaic mutations in the blood of the offspring and were therefore likely to have occurred postzygotically. Furthermore, genome-wide analysis of "*de novo*" variants in the proband led to the identification of 4 out of 4,081 variants which were also detectable in the blood of one of the parents, implying parental mosaicism as the origin of these variants. Thus, our results show that an important fraction of *de novo* mutations presumed to be germline in fact occurred either postzygotically in the offspring or were inherited as a consequence of low-level mosaicism in one of the parent

Introduction

In humans, DNA replication is estimated to entail one error every 10^8 base pairs, giving rise to 30 to 100 genome-wide *de novo* mutations in each new generation.[1–3] While neutral or benign *de novo* point mutations contribute to normal genetic variation, single detrimental *de novo* mutations have been established to cause a number of rare developmental disorders[4–6] and are increasingly recognized as a major contributor to common sporadic disorders, such as intellectual disability (ID) and autism.[7,8] *De novo* mutations are thought to occur predominantly in the egg or sperm cell, resulting in an embryo with a constitutive mutation. However, *de novo* mutations may also appear postzygotically, leading to mosaicism in the embryo, a state in which two or more genetically distinct cell populations in an individual develop from a single fertilized egg.

Several reports have shown a high frequency of mosaicism for copy number variations from cleavage stage embryos[9] to fully differentiated tissues.[10–12] Similarly, there is increasing evidence for a high prevalence of mosaicism for single nucleotide variants (SNVs) as a result of mutations appearing from early embryogenesis onwards[13,14] and throughout adult life.[15,16] Currently, postzygotic *de novo* mutations receive growing attention in developmental diseases.[17–19] The timing of the event plays a key role in the clinical phenotype by determining not only the proportion of affected cells in the organism, but also the type of tissues involved.[18] Despite its pervasiveness, however, the true extent of mosaicism for SNVs remains unclear. This is largely a consequence of the technological limitations to accurately detect these mutations; on one side, mutations with low levels of mosaicism are often below the threshold of sensitivity and specificity for automated and systematic detection of traditional sequencing methods,[20] and on the other hand, mutations with a higher percentage of affected cells are easily detected by traditional sequencing methods but remain technically challenging to differentiate from germline *de novo* mutations. Indeed, to discriminate postzygotic from germline *de novo* mutations by sequencing DNA, it is crucial to distinguish biologically relevant allele imbalances from technical artifacts.

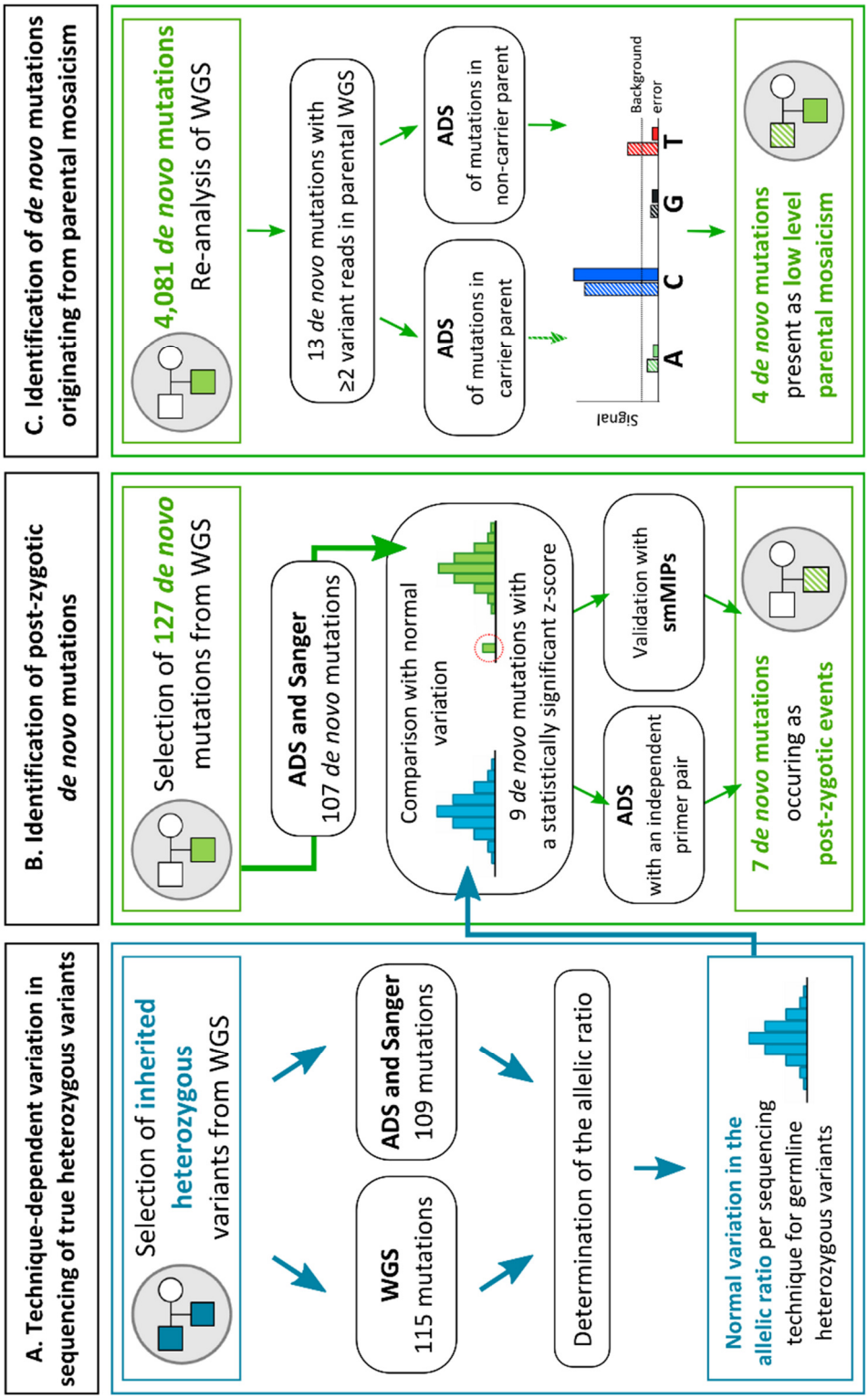
To gain insight into the frequency of postzygotic events among *de novo* mutations, we performed a systematic evaluation of *de novo* mutations identified by trio-based whole genome sequencing (WGS) of 50 individuals with severe ID and their parents. Previous analysis of WGS data from this cohort recently pointed to germline *de novo* mutations as the major cause of ID in the affected individuals.[21] Additionally, these data indicated the presence of *de novo* mutations of somatic origin.[21] By systematic assessment of allelic ratios using various sequencing techniques, we here show that a proportion of previously reported *de novo* mutations do not occur during gametogenesis but arise, in fact, as postzygotic events in the proband or are present as low-level somatic mutations in one of the parents.

Results

Determining the technical variation for WGS, ADS and Sanger sequencing

To gain insight into the sensitivity of WGS, ADS and Sanger sequencing, we re-sequenced two different sets of inherited germline mutations, as proxy for true heterozygosity (Figure 1A and Supplementary Figure S2). We subsequently determined the distribution of the allelic ratios per technology (Supplementary Table S1 and Supplementary Figure S3). With an allele ratio of 48.2±4.4% (average ±standard deviation), ADS showed to be the most precise technique for the identification of true heterozygosity. In comparison, WGS showed an allelic ratio of 50.5±8.9% and Sanger sequencing of 51.4±8.7% (Supplementary Table S2). Based on the obtained distributions for the allelic ratio, we determined that *de novo* mutations with an allelic ratio below 32.8% for WGS, 39.3% for ADS and 33.9% for Sanger sequencing have a statistically significant deviation from the expected ratio for true heterozygous mutations and may, as such, reflect mosaic mutations.

Figure 1. Workflow for the detection of mosaic mutations among a subset of apparently *de novo* mutations. (Part A) Assessment of technique-dependent variation in sequencing of two groups of heterozygous germline variants (in blue) to determine the distribution of allelic ratios for three different techniques (WGS, ADS and Sanger sequencing). (Part B) Previously identified *de novo* mutations were re-sequenced by ADS and Sanger sequencing to determine the variant ratio. Using the combined z-score, nine putative somatic variations were identified, which were validated by ADS with a second independent primer pair and smMIPs. Seven of nine mutations were confirmed to deviate in allelic ratio suggesting a non-germline event. (Part C) Identification of *de novo* mutations originating from parental mosaicism. Of 4,081 high confidence *de novo* mutations identified by WGS, thirteen variants were identified to have two or more variant reads in parental DNA. Using ADS data from the non-carrier parent to correct for the background sequencing error, four mutations appearing as *de novo* in the child were identified as low-level mosaicism in one of the parents. ▶


A. Technique-dependent variation in sequencing of true heterozygous variants
B. Identification of post-zygotic de novo mutations
C. Identification of de novo mutations originating from parental mosaicism
Selection of inherited heterozygous variants from WGS
WGS
115 mutations
ADS and Sanger
109 mutations
Determination of the allelic ratio
Normal variation in the allelic ratio per sequencing technique for germline heterozygous variants
Selection of 127 de novo mutations from WGS
ADS and Sanger
107 de novo mutations
Comparison with normal variation
9 de novo mutations with a statistically significant z-score
ADS
with an independent primer pair
Validation with smMIPs
7 de novo mutations occuring as post-zygotic events
4,081 de novo mutations
Re-analysis of WGS
13 de novo mutations with ≥2 variant reads in parental WGS
ADS
of mutations in carrier parent
ADS
of mutations in non-carrier parent
Signal
A
C
G
T
Background error
4 de novo mutations present as low level parental mosaicism

Identification of postzygotic de novo mutations in probands

Our next objective was to determine the proportion of postzygotic events among a subset of *de novo* mutations in our cohort. For this, we studied a pre-defined set of 107 *de novo* mutations using WGS, ADS and Sanger sequencing (Figure 1B).[21] Similar to the inherited variants, allelic ratios for each mutation were determined for each sequencing technique. After calculation of the mean allelic ratio across the three sequencing techniques, nine *de novo* mutations showed a statistically significant deviation from the expected ratio for true germline heterozygosity (Supplementary Figure S4 and Supplementary Figure S5).

To exclude technical artifacts resulting from biased allele amplification during PCR which would thereby falsely suggest the presence of mosaicism, we generated a second independent amplicon with different PCR primers to re-sequence all nine mutations by ADS (Table 1, Supplementary Table S1 and Supplementary Table S3). This analysis confirmed a statistically significant deviation in the allelic ratio for eight out of nine *de novo* mutations. Of note, three of these mutations had been previously reported as possible mosaic mutations.[21]

To validate these findings with an independent test, we set out to sequence the eight candidate mosaic mutations using smMIPs for increased depth and accuracy. By sequencing germline mutations within the same assay, we first established for this technique the average and standard deviation of the allelic ratio for true heterozygosity, which was shown to be 47.1±3.3%. Unique smMIPs could be designed for all but one candidate mosaic event, located in an intron of *SETBP1* (MIM 611060). The remaining seven mutations were tested and confirmed to be present as mosaic events with allelic ratios between 20.8 and 29.7%. Translating these allelic ratios into percentages of cells carrying the mutation predicts that the mutations must be present in 41.6 to 59.4% of the cells in blood. Thus, our results indicate that at least seven of 107 (6.5%) *de novo* mutations detected in our cohort did not occur in the germline of the parent, but arose postzygotically in the offspring.

Parental mosaicism as a source of seemingly de novo mutations

Gonadal mosaicism in a healthy parent can lead to the transmission of disease-causing mutations and recurrence of disorders with seemingly *de novo* ocurrence.[24] In some cases, mosaicism may not be restricted to the germ cells; it was recently shown that healthy individuals with gonadal mosaicism for disease-causing CNVs, revealed by recurrence of the disease in the offspring, carried low levels of mosaicism for this CNV in blood.[25] Following this idea, we aimed to determine whether any of the seemingly germline *de novo* events in our cohort of 50 probands had actually occurred as somatic mutations in one of the parents (Figure 1C). For this, we re-analyzed all 4,081 high confidence *de novo* mutations

Gene name	Mutation at gDNA level (hg19)	Location	Predicted mutation at cDNA level	Predicted protein substitution	p-value*	Average %
KANSL2	chr12:49072911C>A	Exon 4	NM_017822.3: c.453G>T	p.G151=	6.94E-21	20.8
CREBL2	chr12:12788868G>C	Exon 2	NM_001310.2: c.173G>C	p.R58P	6.40E-19	21.0
PIAS1	chr15:68468014T>A	Exon 10	NM_016166.1: c.1209T>A	p. D403E	1.84E-18	22.9
PNKP	chr19:50367525C>T	Intron 5	NM_007254.3: c.579-32G>A	N/A	7.05E-17	22.7
HIVEP2	chr6:143092683C>T	Exon 5	NM_006734.3: c.3193G>A	p.A1065T	2.20E-14	25.2
DPYD	chr1:97588236C>T	Intron 21	NM_000110.3:c.2623-24048G>A	N/A	3.17E-10	29.7
NEK1	chr4:170359295T>G	Exon 27	NM_001199397.1:c.2703A>C	p.K901N	3.67E-08	29.4

Table 1. *De novo* mutations occurring as post-zygotic events in the offspring. *p-values were corrected by Benjamini-Hochberg for multiple testing; The level of the mutation was calculated by averaging the variant ratio for each mutant from all sequencing methods. N/A, Not applicable.

previously detected by WGS in the probands and selected those *de novo* mutations in which two or more variant reads could be detected in the raw sequence data in one of the respective parents. Thirteen such mutations were identified but two could not be amplified by PCR and were excluded from further analysis. Sequencing by ADS was performed on the remaining 11 mutations to determine whether we could detect the variant in DNA from the carrier parent. After stringent correction for the background sequencing error, four of these mutations were confirmed to be present in blood of one of the parents. These low-level parental mosaic mutations show an average allelic ratio of 3.54% (range 0.22 to 6.15%; Table 2 and Supplementary Table S4). Of note, these low-level parental mosaic mutations, of which three were transmitted by the father and one by the mother, were not detected in the parental DNA by Sanger sequencing (Supplementary Figure S6).

Modeling the effect of sequence coverage on the detection of mosaic mutations

Evidently, sufficient sequencing coverage is required to reliably identify mosaic mutations. To investigate the impact of coverage on the detection of mosaic mutations, we modeled the probability to detect both postzygotic mutations in a proband as well as low-level parental mosaicism given different sequencing coverage.

The detection of postzygotic *de novo* mutations requires two essential steps: calling of the variant in the proband and identification of a significant deviation of the allelic distribution. Modeling under the assumption that ≥5 variant reads are required for variant calling and that these constitute ≥5% of the total number of sequence reads indicates that at least 100-fold coverage is required to call 90% of mosaic variants with an allelic ratio equal to 10% or higher (Supplementary Figure S7). Increased sequencing coverage decreases the standard deviation in the allele ratio, which reduces technical variation (Supplementary Figure S8) and allows for better discrimination between true heterozygosity and mosaicism. Provided that a postzygotic mutation is called, we also modeled the deviation in the allelic ratio of a mosaic variant required to reliably distinguish it from a heterozygous variant (Supplementary Figure S9). Our model indicates that at least 100-fold coverage is required to distinguish mosaic mutations with allelic ratios <40% from germline mutations with 95% probability.

The analysis for parental mosaicism for *de novo* mutations identified in a proband requires a different approach; the identification of parental mosaicism for a seemingly *de novo* mutation in the offspring is guided by the presence of the variant in the proband. As a consequence, the only requirement for the identification of parental mosaicism is to distinguish the variant reads in the parent from the background sequencing error at the respective genomic location.

Genomic location	Gene	Gene location	Origin	Total reads (ADS)	Variant reads (%)	p-value*
chr13:78303535 A>T	*SLAIN1*	Intron	Father	31,470	6.15	<0.001
chr18:25210178 C>T	-	Intergenic	Father	34,149	2.56	<0.001
chr5:11327458 C>T	*CTNND2*	Intron	Mother	12,754	5.25	<0.001
chr5:147855052 G>A	*HTR4*	Intron	Father	20,927	0.22	<0.05

Table 2. *De novo* mutations in the offspring originating from parental mosaicism. *p-values were corrected for multiple testing by Bonferroni correction.

Under the assumption that 2 variant reads in the parent are sufficient for this, we modeled the coverage required to identify low-level parental mosaicism (Supplementary Figure S10) which shows that at least 140-fold coverage is needed to detect low-level mosaicism of ≥5% with ≥95% probability.

Discussion

The aim of our study was to investigate the presence of non-germline events among *de novo* mutations. Our results show that 6.5% of a subset of *de novo* mutations (seven out of 107) were present as mosaic mutations in blood of the proband, strongly suggestive of a postzygotic origin. Extrapolating our results to published genome-wide *de novo* mutation rates[3,21] suggests that each individual carries at least 2 to 7 *de novo* mutations of postzygotic origin. Additionally, from a group of 4,081 mutations presumed to be *de novo* in the offspring, we detected four mutations which were in fact inherited from one of the parents in whom the mutation was present as a low-level mosaic mutation. Although this represents only 0.1% of all high quality *de novo* mutations, parental mosaicism for a seemingly *de novo* mutation in the offspring was observed in four out of 50 trios. Based on the stringent criteria which were used to validate variants as mosaics and our modeling data, we anticipate that our results are likely an underestimation of the true number of mosaic mutations present in blood.

Our initial selection of potential mosaic variants was based on results obtained with relatively high coverage WGS (80-fold). We show that, for trio-based WGS, 80-fold sequencing coverage is sufficient to identify postzygotic events among *de novo* mutations. However, statistical modeling of the probability to detect mosaicism given various sequencing depths showed that, with this coverage, there is only an 80% probability of obtaining sufficient reads to identify mosaicism present in ≥10% of the alleles (corresponding to ≥20% of the cells studied, Supplementary Figure S7). Similarly, with this coverage we are

only able to reliably distinguish somatic events with allelic ratios below 39% from germline mutations (Supplementary Figure S9). This suggests that postzygotic variants in the proband with allelic ratios at either extreme may have gone unidentified in our study. On the other hand, the probability of obtaining at least 2 sequence reads to identify a ≥5% parental mosaicism is only 78% with 80-fold sequencing coverage, suggesting that the identification of these mutations can also be optimized by providing higher sequencing coverage (Supplementary Figure S10). Indeed, the low-level parental mosaic variants identified in our study had a significantly higher sequencing depth in the carrier parent than the other *de novo* or postzygotic mutations studied (Supplementary Figure S11). Our results and statistical modeling highlight the importance of high sequencing coverage in the design of trio-based WGS studies. Currently, most WGS studies are performed at 30-fold coverage.[13,26,27] If we assume that sequence quality is comparable to that of our study, this entails that less than 20% of mosaic variants with an allelic ratio between 10 and 33% can be identified with 30-fold sequencing coverage. Additionally, at this sequencing coverage, only mosaic mutations with an allele ratio below 35% can be reliably distinguished from true heterozygous variants. Furthermore, our modeling suggests that there is less than 20% probability of identifying parental mosaicism with an allelic ratio of less than 5% with WGS at 30-fold coverage. Given these results, our findings underline the need for increased sequencing coverage in WGS for the accurate identification of mosaicism.

Despite the aforementioned limitations, we show that WGS is a powerful method for genome-wide discovery of mosaic events. In this study, we used three additional techniques to confirm mosaicism of SNVs. After identification of *de novo* mutations by WGS, their status as postzygotic events was first evaluated by ADS and Sanger sequencing. A limitation of both these techniques is that they may show an allelic imbalance due to biased amplification of one allele over the other.[28] For the most part, significant deviations in the allele ratio secondary to technical artifacts observed in Sanger and ADS were method–specific, rather than reproducible PCR artifacts (Supplementary Figure S12). We have attempted to remedy this problem by the use of smMIPs, which provide targeted high sequence coverage and the ability to identify individual captured molecules,[23] thus preventing any deviations in the allele ratio due to PCR amplification bias.

The presence of parental gonosomal mosaicism as the cause of a sporadic disorder in a family places the subsequent offspring at higher risk for recurrence of the disease than when the mutation is caused by a germline *de novo* mutation.[29] Considering this, the presence of parental mosaicism in four out of 50 individuals of our cohort stresses the importance of a thorough follow-up in families affected by a disorder due to a *de novo* mutation.[30] Notably, the lower limit of detection by Sanger sequencing has been reported to be close to only 10%,[25] whereas the highest level of parental mosaicism here detected was only 6.15% and could not be identified by Sanger sequencing (Supplementary Figure S6). Since Sanger sequencing is commonly used in diagnostics, parental

mosaicism below the threshold of detection of this method could account for recurrence of *de novo* disorders within families[24,31] and explain unsolved pedigrees with an apparently recessive inheritance of disorders otherwise known to be dominant.[32] Under these circumstances, high coverage NGS should be favored over Sanger sequencing for the detection of low-level parental mosaicism and may even be warranted as a standard follow-up test for each pathogenic *de novo* mutation. Related to this, the frequent detection of mosaic events may partially explain the occurrence of known dominant pathogenic mutations within large-scale variant databases of healthy individuals such as Exome Variant Server. This point needs to be taken into account when these databases are used for clinical interpretation of possible pathogenic mutations. Also, previous studies have shown that certain mutations found as true heterozygous events in one tissue may be detected at low levels or completely absent in another.[33] Clearly, further studies of mosaic mutations and their impact on phenotypic variation requires an in-depth analysis of different tissues.

In summary, our results show that a proportion of *de novo* mutations presumed to be germline actually either occurred postzygotically in the offspring or were inherited from low-level mosaicism in one of the parents. This indicates that *de novo* mutations do not arise solely during gametogenesis but also as postzygotic mutations, suggesting that our genomes may be much more dynamic than previously considered. As the contribution of *de novo* mutations to human disease becomes increasingly apparent, this conclusion may well have clinical implications. Pathogenic variants in mosaic state require particular attention as to their detection using sequencing methods. Furthermore, their influence on the risk of recurrence of a disease underlines the importance of the identification of mosaicism to offer accurate genetic counseling in sporadic disorders caused by *de novo* mutations.

Materials and methods

Defining a set of de novo mutations from whole genome sequencing of parent-proband trios

This study was performed in accordance with the ethical standards of the Medical Ethics Committee of the Radboud university medical center. All participants or their legal representatives gave informed consent. WGS of 50 parent-proband trios and subsequent *de novo* mutation detection was performed as described previously.[21] Briefly, trio-based WGS was performed by Complete Genomics (CG) at 80-fold coverage. Sequence reads were mapped to the reference genome (GRChr37) and variants were called using CG software v.2.4. *De novo* mutations were called using CG's cgatools calldiff program, which detects the differences between the genotypes of two samples and assigns a somatic score based on sequencing quality and paired sample comparison. Mutations with scores for offspring versus both parents ≥5 were called as high confidence *de novo* mutations (for which a total of 4,081 was detected in the 50 trios). The original report identified a set of 127 *de novo* mutations, affecting either genome-wide coding sequence or specifically the non-coding sequence of known ID genes.[21] This set served as the starting point for the current study.

Sequencing methods used to assess the postzygotic state of de novo mutations

PCR amplicons for amplicon-based deep sequencing (ADS) and Sanger sequencing were generated using standard PCR protocols. ADS was performed on an Ion Torrent PGM sequencer (Life Technologies, Carlsbad, CA, USA) as described previously.[21] In brief, raw sequencing reads were mapped to the reference genome using the BWA software package and the alignment files were then analyzed in the Integrative Genomics Viewer (IGV).[22] For Sanger sequencing, PCR products were sequenced after enzymatic clean-up.

Sequencing using single molecule molecular inversion probes (smMIPs) was done following previously published protocols.[23] Briefly, smMIP probes targeting the selected *de novo* mutation and a total of 112 bp surrounding sequence were designed in-house and ordered from Integrated DNA Technologies (IDT, Leuven, Belgium). The smMIPs were pooled and phosphorylated, after which the genomic regions of interest were captured with the probes and amplified. Sequencing was performed on the NextSeq 500 desktop sequencer (Illumina, San Diego, CA, USA) and the reads were aligned using our in-house bioinformatics pipeline for molecular inversion probe analysis. Through the use of molecular barcodes, we were able to remove PCR duplicates. Read counts for the positions of interest were extracted from the alignment files through IGV.

Assessment of the allelic ratio distribution of true heterozygous variants

To define the parameters of technical variation in WGS, ADS and Sanger sequencing, we determined for each technology the allelic ratio of inherited SNVs, as proxy for true heterozygous mutations. The allelic ratio was defined as the proportion of variant reads from the total number of sequencing reads covering a given base pair and is expressed here as a percentage. The distribution of the allelic ratio for true heterozygous variants in WGS data was established by determining the allelic ratio of 115 inherited SNVs from WGS data of a single individual (coding, synonymous variants absent from dbSNP138 or present at a frequency below 1.5%). To minimize the risk of false-positive variant calls, a second independent set of 109 inherited SNVs was used to determine the distribution of the allelic ratio in ADS and Sanger sequencing. This set was randomly selected from a larger set of 442 rare, coding variants inherited from either parent in 10 probands and variants were selected to have a coverage ≥20-fold in WGS, with a percentage of variant reads between 40 and 60%. Variants on the X chromosome and/or located in established disease genes were excluded. For ADS experiments, after mapping with BWA, variants were visualized using IGV and allelic ratios were determined by assessing the number of total reads and each respective base at this position. For Sanger sequencing, the chromatogram trace files were visualized using Vector NTI (Life Technologies, Carlsbad, CA, USA) and intensities per dye per variant base were used to calculate the allelic ratio.

Identification of postzygotic events in probands

A set of 127 *de novo* mutations identified by WGS was re-sequenced by ADS and Sanger sequencing. For 107 of these variants (84%), allelic ratios could be determined for all three sequencing techniques. Using the distribution of the allelic ratio in true heterozygosity as a reference, we calculated the individual z-score per method for each mutation. Z-scores were calculated using the values from sequencing heterozygous variants with each sequencing method as a reference and are defined as the difference between the value of the allelic ratio and the mean allelic ratio for heterozygous variants on that sequencing technique, divided by the standard deviation. Subsequently, these scores were brought together into a single combined z-score for each *de novo* mutation by summing the individual z-scores and dividing this total by the square root of the number of scores. The critical value for statistical significance was established at 0.05 after Benjamini-Hochberg correction for multiple testing. To exclude amplification bias as the cause of a deviation in the allele ratio, *de novo* SNVs with a statistically significant combined z-score were re-sequenced by ADS using a second independent primer pair. Finally, smMIPs were used as an independent technique to validate the presence of these variants as mosaic mutations, using a set of 7 heterozygous mutations as a reference.

Identification of parental mosaicism in whole genome sequencing data

To detect low-level parental mosaicism for SNVs mimicking germline *de novo* mutations in the child, we re-analyzed the WGS data of the 50 parent-offspring trios. To this end, we used all 4,081 high confidence candidate *de novo* mutations identified in the probands, as these have previously been shown to have a *de novo* validation rate of 78%.[21] We then filtered for *de novo* variants for which at least two reads carrying the same mutation in the raw sequencing data were found in either one of the parents. Sequencing of the position of interest by ADS in the DNA of the transmitting parent was performed to validate parental mosaicism for the remaining 11 mutations. The position-specific sequencing error rate was established by sequencing the same position by ADS in the DNA of the non-transmitting parent in an independent sequencing run to avoid any contamination or barcode bleed-through. Then, the fraction of reads showing a non-reference allele at the corresponding base pair was calculated. The presence of the variant as a mosaic mutation in the transmitting parent was confirmed if the proportion of variant reads for the position and nucleotide of interest was significantly higher than the sequencing error for that base-pair position established from the non-transmitting parent.

Computational modeling of sequencing coverage for the identification of mosaicism

To assess the ability of identifying mosaic variants from sequencing data, we simulated the effect of sequencing coverage on variant identification for different levels of mosaicism. To distinguish low-level mosaicism from sequencing artifacts, we assumed that automated variant-calling algorithms require the variant to be present in ≥5 sequencing reads and constitute ≥5% of the total number of reads at the position of interest. A binomial distribution was used to calculate the probability of reaching both these requirements for different depths of coverage and various levels of mosaicism. Assuming that a mosaic variant has been identified, we also modeled the deviation of the allelic ratio from 50% (representing true heterozygosity) which is necessary to distinguish a mosaic from a germline variant. Reads for heterozygous variants at different sequencing depths were simulated (n=10,000) based on a binomial distribution. The standard deviation of this distribution was calculated and we assessed the level of mosaicism which could be reliably distinguished from a heterozygous variant for different thresholds of significance. Lastly, we determined the sequence coverage which would be required to identify low-level parental mosaicism. In this case, the position of interest is readily identified due to the offspring presenting an apparently *de novo* mutation at this position. For this, we considered that at least 2 variant reads are sufficient to distinguish the variant from background sequencing error. Finally, we applied a binomial model for different sequencing

depths and levels of mosaicism to calculate the probability of obtaining 2 variant reads in the sequencing data.

Web resources

Integrated Genomics Viewer: http://www.broadinstitute.org/igv/
Exome Variant Server: http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man: http://www.omim.org/

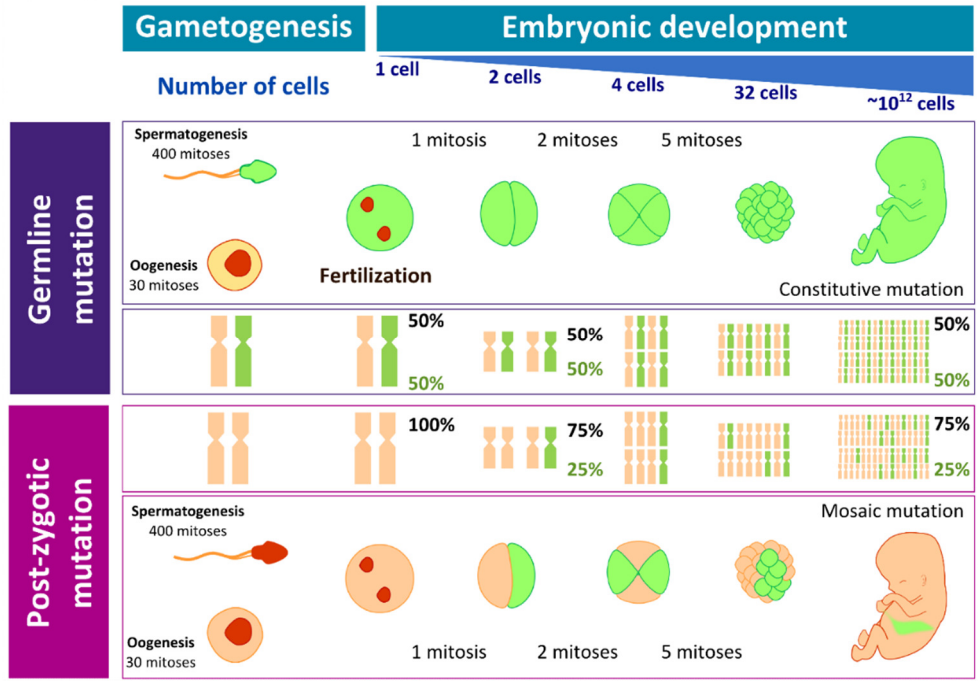
References

1. Nachman, M. W. M. & Crowell, S. S. L. Estimate of the Mutation Rate per Nucleotide in Humans. *Genet* **156,** 297–304 (2000).
2. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).
3. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43,** 712–714 (2011).
4. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* **42,** 483–485 (2010).
5. Rivière, J.-B. *et al.* De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. *Nat Genet* **44,** 440–4, S1-2 (2012).
6. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42,** 790–793 (2010).
7. Vissers, L. E. L. M. *et al.* A de novo paradigm for mental retardation. *Nat Genet* **42,** 1109–1112 (2010).
8. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43,** 585–589 (2011).
9. Vanneste, E. *et al.* Chromosome instability is common in human cleavage-stage embryos. *Nat Med* **15,** 577–583 (2009).
10. O'Huallachain, M. *et al.* Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci U S A* **109,** 18018–18023 (2012).
11. McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science (80-)* **342,** 632–7 (2013).
12. Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492,** 438–442 (2012).
13. Dal, G. M. *et al.* Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J Med Genet* **51,** 455–9 (2014).
14. Huang, A. Y. *et al.* Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res* **24,** 1311–1327 (2014).
15. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20,** 1472–1478 (2014).
16. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N Engl J Med* **371,** 2488–2498 (2014).
17. Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* **365,** 611–9 (2011).
18. Shirley, M. D. *et al.* Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. *N Engl J Med* **368,** 1971–9 (2013).
19. Kurek, K. C. *et al.* Somatic Mosaic Activating Mutations in PIK3CA Cause CLOVES Syndrome. *Am J Hum Genet* **90,** 1108–1115 (2012).
20. Rohlin, A. *et al.* Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. *Hum Mutat* (2009). doi:10.1002/humu.20980
21. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).
22. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* (2011). doi:10.1038/nbt.1754
23. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* **23,** 843–854 (2013).

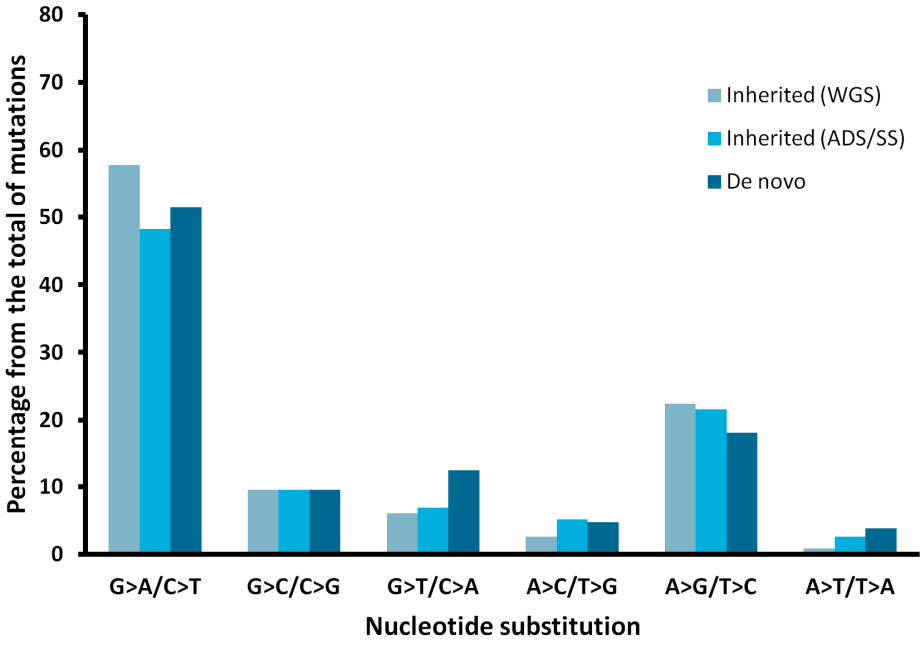
24. Natacci, F. *et al.* Germline mosaicism in achondroplasia detected in sperm DNA of the father of three affected sibs. *Am J Med Genet Part A* (2008). doi:10.1002/ajmg.a.32228
25. Campbell, I. M. *et al.* Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. *Am J Hum Genet* **95,** 173–182 (2014).
26. Petersen, B.-S. *et al.* Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease.
27. Nemirovsky, S. I. *et al.* Whole Genome Sequencing Reveals a De Novo SHANK3 Mutation in Familial Autism Spectrum Disorder. *PLoS One* **10,** e0116358 (2015).
28. Veal, C. D. *et al.* A mechanistic basis for amplification differences between samples and between genome regions. *BMC Genomics* **13,** 455 (2012).
29. Campbell, I. M. *et al.* Parent of origin, mosaicism, and recurrence risk: Probabilistic modeling explains the broken symmetry of transmission genetics. *Am J Hum Genet* **95,** 345–359 (2014).
30. Faivre, L. *et al.* Recurrence of SOX2 anophthalmia syndrome with gonosomal mosaicism in a phenotypically normal mother. *Am J Med Genet Part A* **140A,** 636–639 (2006).
31. Elalaoui, S. C. *et al.* Germinal mosaicism in Noonan syndrome: A family with two affected siblings of normal parents. *Am J Med Genet Part A* (2010). doi:10.1002/ajmg.a.33685
32. Schinzel, A. & Giedion, A. A syndrome of severe midface retraction, multiple skull anomalies, clubfeet, and cardiac and renal malformations in sibs. *Am J Med Genet* **1,** 361–375 (1978).
33. Huisman, S. A., Redeker, E. J. W., Maas, S. M., Mannens, M. M. & Hennekam, R. C. M. High rate of mosaicism in individuals with Cornelia de Lange syndrome. *J Med Genet* **50,** 339–344 (2013).

Supplementary data

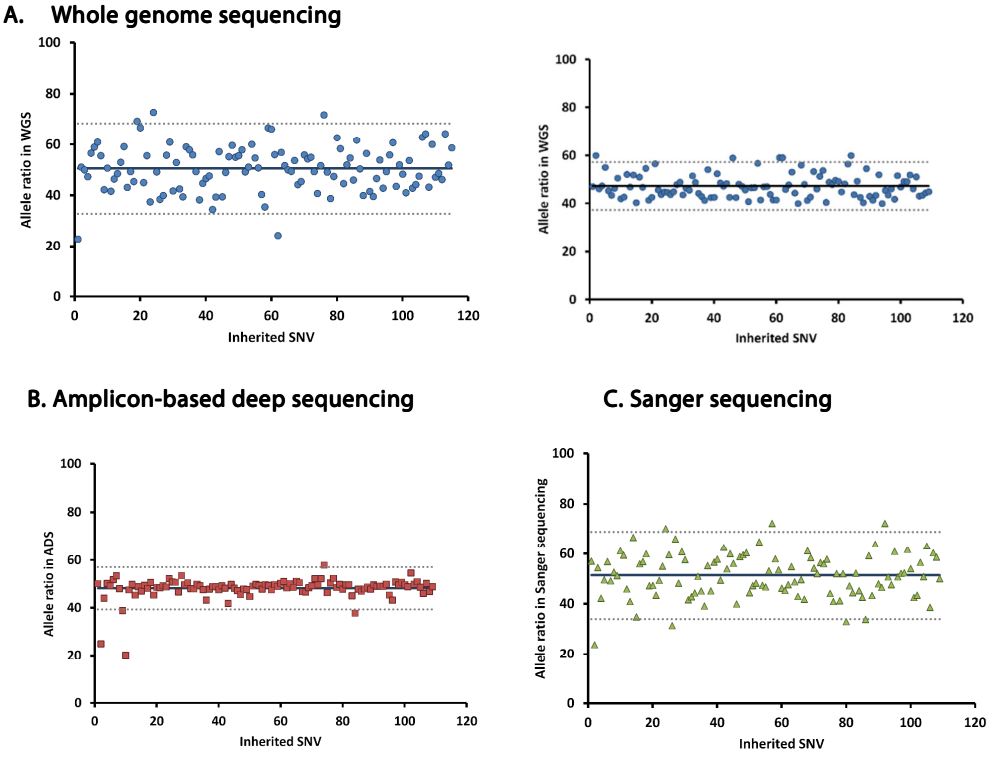
Supplementary Figures



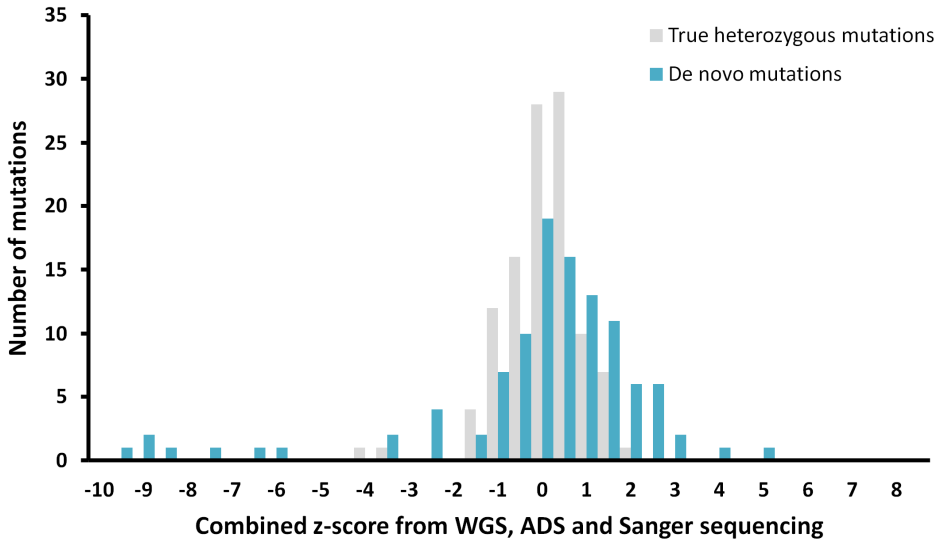
Supplementary Figure S1. Germline *de novo* mutations are present in all cells and are therefore truly heterozygous, with an equal distribution of the wild-type and the mutated allele (50:50, see top panel "germline mutation"). Somatic *de novo* mutations are not present in all cells of an organism; some cells will carry the mutation while others will not, leading to an unbalanced distribution of the mutated allele (*e.g.* 80:20, see bottom panel "post-zygotic mutation"). This unbalanced distribution of the mutated allele proper to somatic mutations can be detected by sequencing as a deviation in the allele ratio (*i.e.* the signal corresponding to the mutant allele versus the signal corresponding to the reference allele). For next generation sequencing (NGS) techniques, this is observed as lower number of reads carrying the mutated allele versus reads carrying the reference allele. When using Sanger sequencing, this unbalance is detected as a difference in the intensity of the bases in the chromatogram. However, there may also be an unbalance in the distribution of the alleles as a result of technical variation during sequencing. To identify mutations present in mosaic state, it is necessary to be able to differentiate the deviation in the variant ratio that is a result of technical variation from the deviation in the variant ratio that is secondary to a biological allele unbalance. Cells and chromosomes carrying a mutation are shown in green in the figure.



Supplementary Figure S2. Frequency of nucleotide substitutions per group of studied mutations. The frequency for all nucleotide substitutions was determined for each of the studied groups (115 inherited variants studied by WGS, 109 inherited variants studied by ADS and Sanger sequencing (SS) and 107 out of 109 evaluated *de novo* mutations). Two mutations from the group of *de novo* variants were excluded from this analysis, as they were deletions and could not be classified. By the means of statistical analysis, we determined that the nucleotide substitution frequency is not significantly different between the analyzed groups (Chi square test, df = 6, X^2=6.113, p = 0.41).

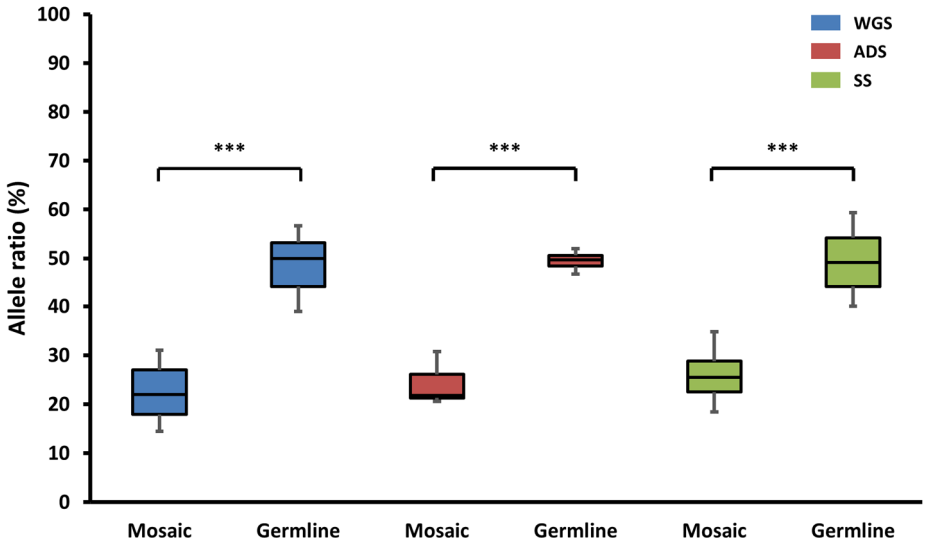


Supplementary Figure S3. Allele ratio of inherited heterozygous variants sequenced using different sequencing techniques. Results include whole genome sequencing (panel A), amplicon-based deep sequencing (panel B) and Sanger sequencing (panel C). For the WGS data, results from 115 variants used to determine the allelic distribution of heterozygous variants are shown in left side of panel A, while the 109 variants used to establish the distribution in ADS and Sanger sequencing are shown on the right side of panel A. The allele ratio was calculated as the percentage of variant reads from the total of reads for NGS technologies and as the intensity of the mutated base in the chromatogram versus the sum of the intensities of the reference and the mutated bases for Sanger sequencing. The mean allele ratio ±2 standard deviations of each technique are also shown (marked by a black line and dotted gray lines, respectively). Refer to Supplementary Table S1 for the raw data.

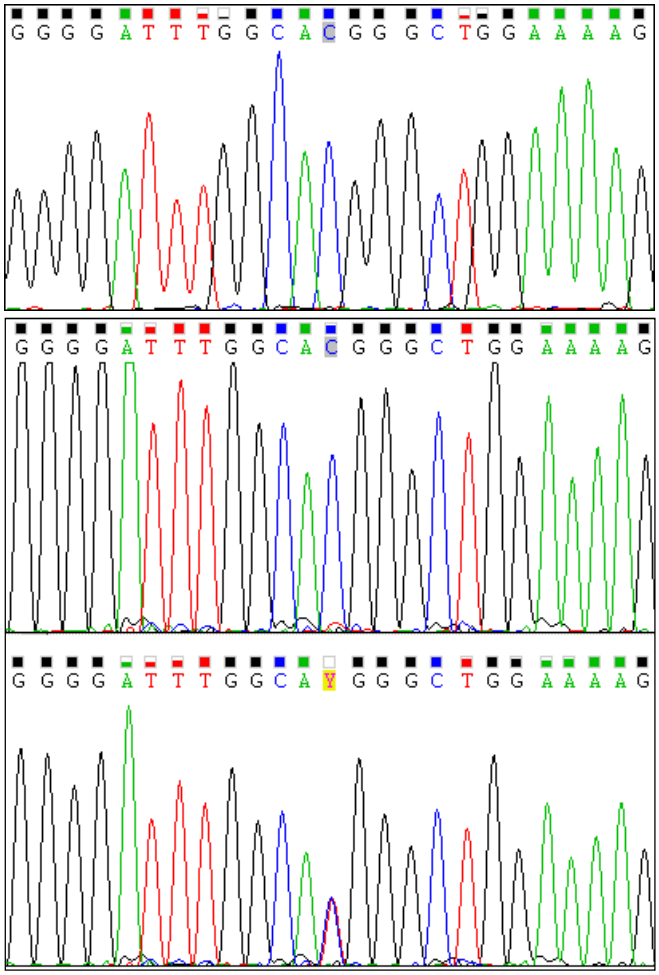


Supplementary Figure S4. Distribution of *de novo* versus true heterozygous variants (*i.e.* inherited heterozygous variants) as a histogram of the number of cases per combined z-score obtained with multiple sequencing techniques (WGS, ADS and Sanger sequencing). On the far left, several outlying *de novo* mutations do not fit the expected distribution for heterozygous variants, suggesting a statistically significant unbalance between the wild-type and the mutant allele. These mutations are present in mosaic state and represent post-zygotic mutations.

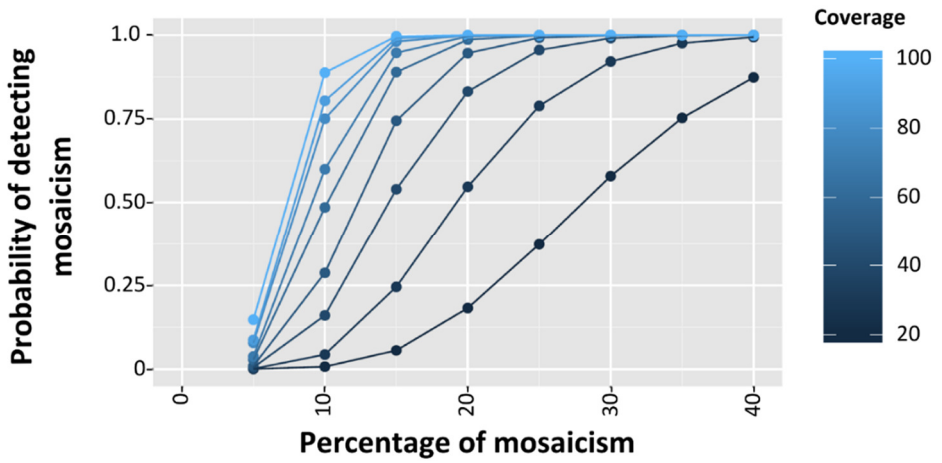
We looked further into the *de novo* mutations found on the extreme right of the histogram (with a z-score of 3.64 and 4.88). Both these mutations are small deletions which have a high combined z-score value due to extreme deviation of the allele ratio higher than the reference ratio only visible in ADS (*i.e.* more than 50% of the total reads in ADS correspond to variant reads). The 7 variants presented as somatic mutations in our study show consistently a low allele ratio in different sequencing techniques and when sequencing was performed using independent primer pairs for amplification.



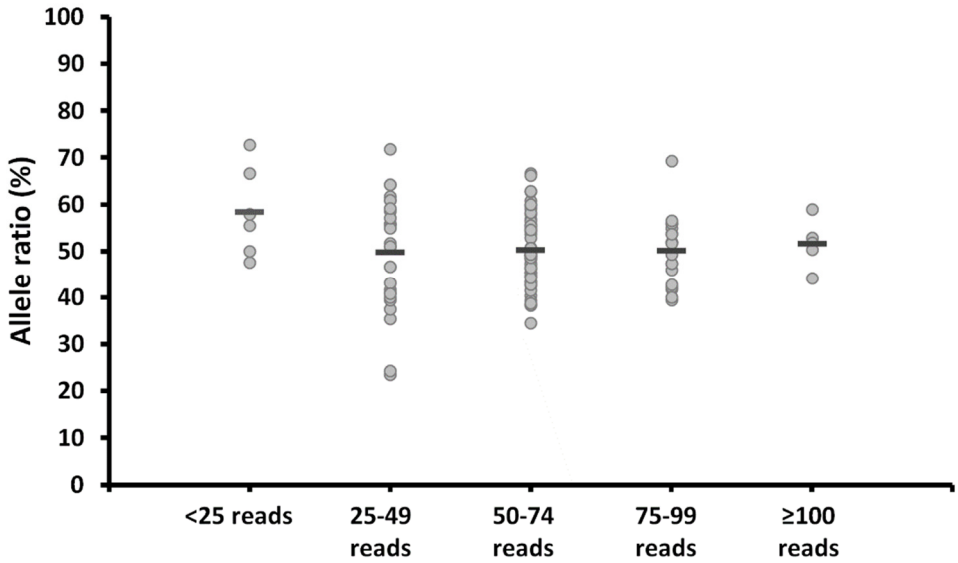
Supplementary Figure S5. Distribution of the allele ratio per sequencing technique in 7 *de novo* mutations identified as post-zygotic events compared to 100 germline *de novo* mutations in the proband. We plot here the median and the 10th, 25th, 75th and 90th percentile for each group. The difference between the variant ratio of postzygotic versus germline mutations was statistically significant for all methods (Student's T-test, *** p<0.001).



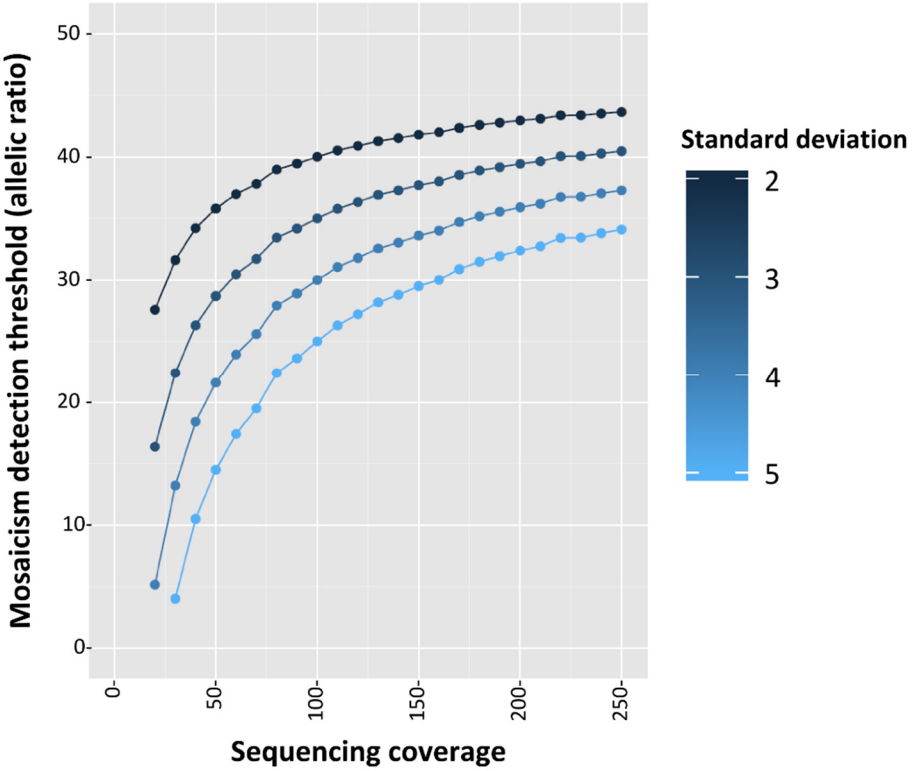
Supplementary Figure S6. Sanger sequencing traces of *de novo* mutation chr5:11327457 C>T in *CTNDD2* in the non-carrier father (top), carrier mother (middle) and proband (bottom). This mutation was originally identified by trio-based WGS and later confirmed in DNA derived from maternal blood by ADS, with a variant ratio of 5.25%. Note that the mutated allele (A, in red) is present in the maternal DNA sample but is indistinguishable from the background traces.



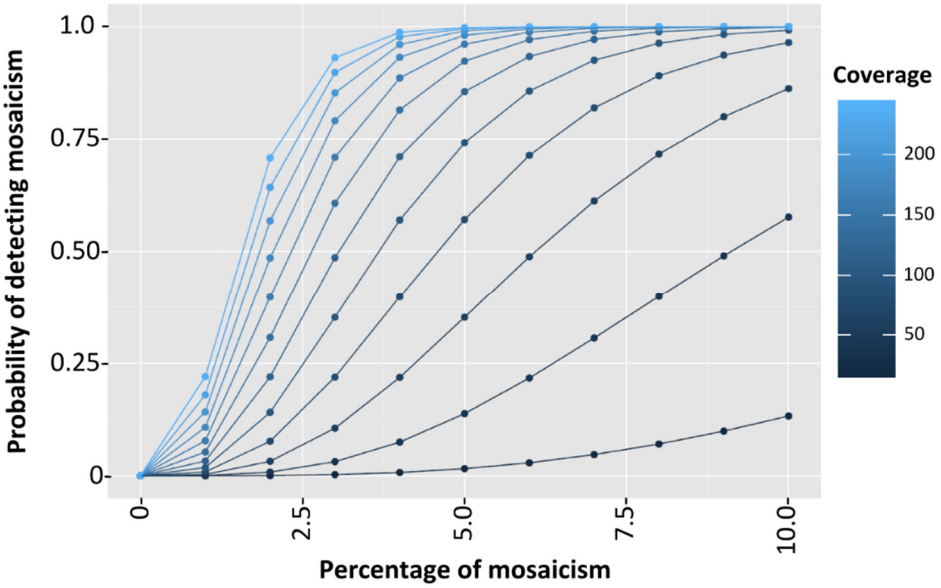
Supplementary Figure S7. Modeling of the probability of identifying variants with different levels of mosaicism given different sequencing coverage. We assumed that automated variant calling algorithms require a variant to be present in at least 5 sequencing reads which constitute at least 5% of the total number of reads at a position. A binomial distribution was used to calculate the probability (*i.e.* power) of reaching both these requirements for different depths of coverage and various levels of mosaicism. The X-axis indicates the percentage of mosaicism (*i.e.* the true allelic ratio). The Y-axis shows the probability of identifying this mosaicism. Each line represents the results for different sequencing coverage according to the legend.



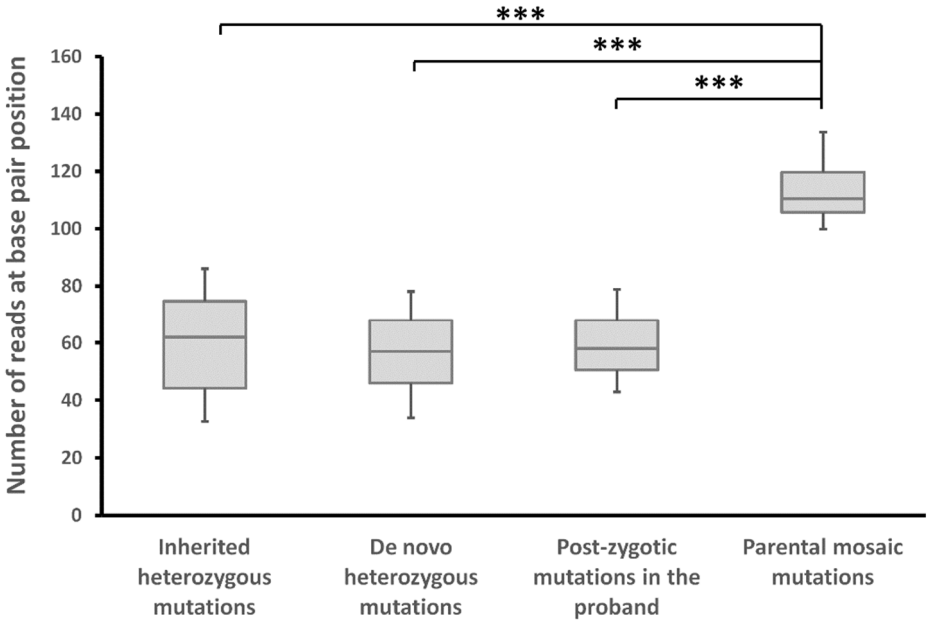
Supplementary Figure S8. Allele ratio per sequencing coverage of 115 inherited heterozygous variants in WGS data. Evaluated variants were classified according to sequencing depth at the base pair position (in bins increasing by 25 sequence reads) and the allele ratio for the variant allele was determined. Each mutation is shown as a circle, with the horizontal bar representing the average allele ratio per category. The groups comprise 6 (<25 reads), 28 (25-49 reads), 56 (50-74 reads), 19 (75 to 99 reads) and 5 variants (≥100 reads). The average allele ratio does not significantly change with increases in sequencing depth above 25-fold coverage (49.8% with 25-49 fold coverage versus 51.7% with ≥100-fold coverage). However, higher sequencing coverage decreases the standard deviation (11.9 with 25-49-fold coverage versus 5.27 with coverage ≥100-fold). This indicates that higher sequencing coverage decreases the technical variation and offers higher sensitivity for the detection of biologically relevant deviations in the variant ratio.



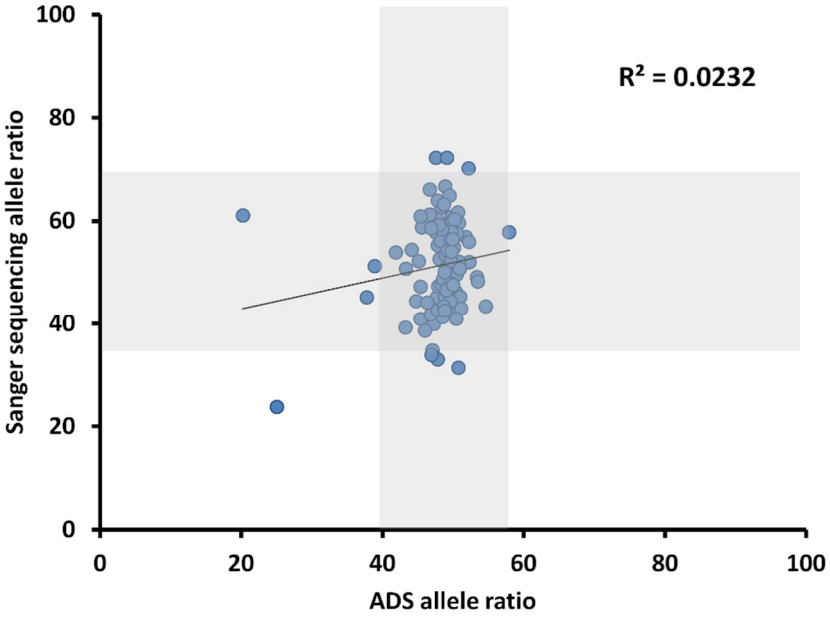
Supplementary Figure S9. Simulation of the level of mosaicism which can be statistically distinguished from a heterozygous variant given different sequencing coverage and thresholds for significance. We simulated reads for heterozygous variants (n=10,000) at different sequencing coverage based on a binomial distribution. From this, we calculated the standard deviation of this distribution and, for different significance thresholds, we assessed the level of mosaicism for which the allelic ratio can reliably be distinguished from the allelic ratio of a heterozygous variant. The X-axis indicates the sequencing coverage, while the Y-axis indicates the level of mosaicism which can be distinguished from a heterozygous variant (note that 10% of the allelic ratio means that 20% of the cells carry the variant). Each line represents the significance threshold as the distance from the average in standard deviations, according to the legend.



Supplementary Figure S10. Modeling of the probability of identifying different levels of mosaicism in at least three reads for different sequencing depths. In this case the position of interest is already identified, as the offspring will have a *de novo* mutation at this base pair. We considered 3 reads showing the mutated allele to be sufficient to distinguish the variant from background sequencing error. We applied a binomial model for different sequencing depths and levels of mosaicism to calculate the probability of obtaining 3 sequencing reads with the variant. The X-axis indicates the percentage of mosaicism as the allelic ratio, while the Y-axis indicates the probability of identifying at least 3 reads. Each line shows the result for different depths of coverage, according to the legend.



Supplementary Figure S11. Sequencing coverage in WGS data per mutation category. Sequencing depth in WGS data for the evaluated mutations are presented per category. The mutations include 115 randomly selected inherited heterozygous mutations (WGS from the proband), 100 *de novo* heterozygous mutations (WGS from the proband), 7 postzygotic *de novo* mutations representative of high-level mosaicism (WGS from the proband) and 4 parental somatic mutations present as low-level mosaicism (WGS from the parent). Plotted here are the median and the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile for each group. The asterisks denote that the difference in sequencing coverage between inherited heterozygous, germline *de novo* variants and postzygotic mutations in the proband and parental mosaic mutations is statistically significant (*** $p < 0.001$, Student's t-test). This suggests that the sequencing coverage required for the detection of *de novo* mutations is lower than the sequencing depth necessary for the detection of low-level parental mosaicism. This finding supports that WGS which would allow for identification of *de novo* mutations may not detect low-levels of parental mosaicism. As a consequence, it is likely that our results are an underestimation of the true extent of *de novo* mutations originating from parental mosaicism.



Supplementary Figure S12. Comparison of the allele ratio obtained for different sequencing techniques in truly heterozygous variants. A group of 109 inherited variants were amplified using the same primer pair and sequenced both by ADS and Sanger sequencing. Each circle in this graph represents one variant, while the gray rectangles highlight the 95% confidence interval for each sequencing method. While there are several variants falling out of the 95% confidence interval for each method, only one SNV shows a statistically significant deviation in the allele ratio both in ADS and Sanger sequencing. This deviation in both sequencing methods may be secondary to biased allele amplification, while deviations observed in a single technique but not reproducible in another may be caused by technical error specific to each sequencing method.

Supplementary Tables

Supplementary Table S1. Raw data for the three groups of variants used in this study. Available at http://www.sciencedirect.com/science/article/pii/S0002929715001949.

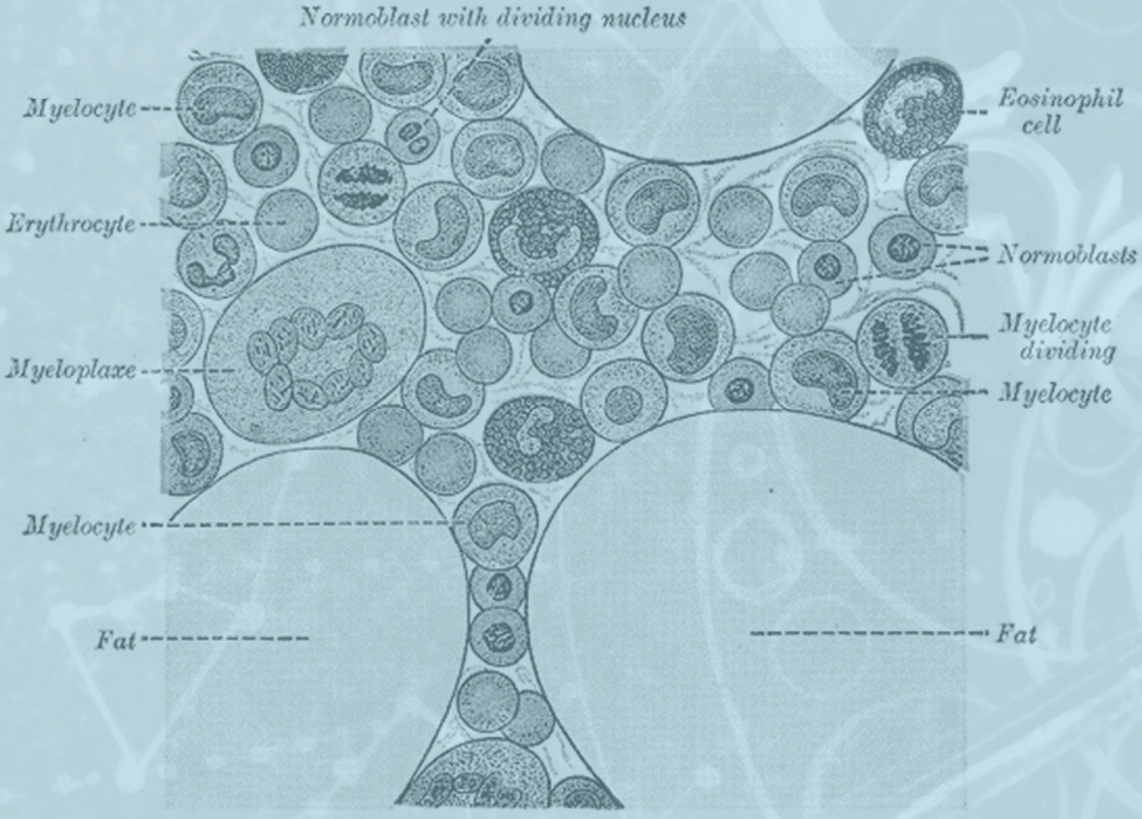
	Whole genome sequencing	Amplicon-based deep sequencing	Sanger sequencing
	Allele ratio %	**Allele ratio %**	**Allele ratio %**
Average	50.5	48.2	51.4
Standard deviation	8.9	4.4	8.7
95% interval	32.8-68.3	39.3-57.0	33.9-68.8
Maximum observed	72.7	57.9	72
Minimum observed	22.9	20.2	24

Supplementary Table S2. Table showing the technical specifications for each sequencing technique, including average variant ratio (expressed as the percentage of variant reads from total reads), standard deviation, 95% range and maximum and minimum observed in the studied germline heterozygous variants. Inherited heterozygous variants were sequenced using Whole Genome Sequencing, amplicon-based deep sequencing and Sanger sequencing.

Supplementary Table S4. Detailed results from WGS and deep sequencing of *de novo* mutations evaluated for parental low-level somatic mosaicism. Available at http://www.sciencedirect.com/science/article/pii/S0002929715001949.

		WGS		Amplicon-based deep sequencing		Sanger sequencing		Amplicon-based deep sequencing (2)		Statistical analysis			Single molecule MIPs	
Gene name	Genomic location (hg19)	mutant %	z-score	mutant %	z-score	mutant %	z-score	mutant %	z-score	Combined z-score	p-value (BH)	Average mutant %	mutant %	z-score
KANSL2	chr12:49072911C>A	21	-3.35	20	-6.32	19	-3.70	19	-6.55	-9.96	6.94E-21	20.8	24.7	-6.85
CREBL2	chr12:12788868G>C	14	-4.14	20	-6.32	31	-2.36	21	-6.09	-9.46	6.40E-19	21.0	19.4	-8.51
PNKP	chr19:50367525C>T	23	-3.13	22	-5.87	25	-2.98	23	-5.64	-8.81	7.05E-17	22.7	20.2	-8.25
PIAS1	chr15:68468014T>A	22	-3.24	22	-5.87	25	-3.08	20	-6.32	-9.25	1.84E-18	22.9	25.9	-6.50
HIVEP2	chr6:143092683C>T	31	-2.22	22	-5.87	31	-2.37	23	-5.64	-8.05	2.20E-14	25.2	19.5	-8.43
NEK1	chr4:170359295T>G	15	-4.06	32	-3.61	40	-1.26	34	-3.16	-6.05	3.67E-08	29.4	25.7	-6.79
DPYD	chr1:97588236C>T	31	-2.22	30	-4.07	27	-2.85	29	-4.29	-6.71	3.17E-10	29.7	31.9	-5.25

Supplementary Table S3. Table showing the detailed results for each sequencing technique (WGS, ADS with two independent primer pairs, SS and smMIPs) per mutation, including average percentage of mutant reads from total reads, technique-specific z-score, combined z-score, p-value corrected for multiple testing using Benjamini-Hochberg correction and the average level of the mutation in blood, estimated from the sequencing results.



Human bone marrow. Highly magnified.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 5:
Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life

Submitted as:

Acuna-Hidalgo R., Sengul H., Steehouwer M., van de Vorst M., Vermeulen S.H., Kiemeney L.A.L.M., Veltman J.A., Gilissen C. & Hoischen A. Somatic driver mutations in age-related clonal hematopoiesis by ultra-sensitive sequencing.

Abstract

Clonal hematopoiesis results from somatic mutations in hematopoietic stem cells, which give an advantage to mutant cells, driving their clonal expansion and potentially leading to leukemia. The acquisition of clonal hematopoiesis-driver mutations (CHDMs) occurs with normal aging and these mutations have been detected in over 10% of individuals ≥65 years.

We aimed to examine the prevalence and characteristics of CHDMs throughout adult life. We developed a targeted re-sequencing assay combining high-throughput with ultra-high sensitivity based on single-molecule molecular inversion probes (smMIPs). Using smMIPs, we screened over 100 loci for CHDMs in over 2,000 blood DNA samples from population controls between 20 and 69 years of age. Loci screened included 40 regions known to drive clonal hematopoiesis when mutated and 64 novel candidate loci. We identified 224 somatic mutations throughout our cohort, of which 216 were coding mutations in known driver genes (*DNMT3A*, *JAK2*, *GNAS*, *TET2* and *ASXL1*), including 196 point mutations and 20 indels. Our assay's improved sensitivity allowed to detect mutations with variant allele frequencies as low as 0.001. CHDMs were identified in over 20% of individuals 60 to 69 years of age and in 3% of individuals 20 to 29 years of age, approximately double the previously reported prevalence despite screening a limited set of loci.

Our findings support the occurrence of clonal hematopoiesis-associated mutations as a widespread mechanism linked with aging, suggesting that mosaicism as a result of clonal evolution of cells harboring somatic mutations is a universal mechanism occurring at all ages in healthy humans.

Introduction

Low level mosaicism resulting from somatic mutations is frequent in healthy tissues,[1] particularly in those with high turnover rates such as blood[2–6] and skin.[7,8] Novel mutations may arise due to failure to repair DNA replication errors[9] or secondary to DNA damage caused by exposure to endogenous and exogenous mutagens.[10] While most somatic mutations are phenotypically silent in the cell in which they arise, some of them can lead to changes in cell behavior. For example, mutations abolishing the function of a gene can be detrimental or even lethal for the cell in which they arise. In contrast, a subset of mutations have the ability to promote cell proliferation and/or survival, granting mutant cells a growth advantage compared to wild-type ones.[11,12] This fitness advantage can allow a single mutant cell to grow into groups of identical daughter cells, which is known as "clonal expansion".[13,14] Mutations driving clonal expansion can arise in all cell types including somatic stem cells, which are characterized by their longevity and continuous division. These two characteristics would allow somatic stem cells to undergo recurring cycles of acquisition of mutations and subsequent clonal expansion, leading to the accumulation and propagation of mutations over time.

A number of recurrent somatic mutations have been implicated in "clonal hematopoiesis", a process in which a mutant hematopoietic stem cell (HSC) expands clonally and contributes to a significant and detectable fraction of circulating blood cells.[2–5] Somatic mutations involved in clonal hematopoiesis are often detected in blood-derived DNA at a variant allelic frequency (VAF) ranging from 0.008 to 0.1, suggesting that between 1.6 and 20% of nucleated cells circulating in blood are derived from mutant HSCs.[2–5] Clonal hematopoiesis driver mutations (CHDMs) most often disrupt genes such as *DNMT3A*, *TET2* and *ASXL1,* which are associated with blood disorders like myelodysplasia and leukemia.[2–4] Because of this link, the acquisition of CHDMs has been suggested to represent the earliest phase in the development of hematologic malignancies.[15] Indeed, clonal hematopoiesis of indeterminate potential, defined by the presence of

CHDMs with a VAF ≥0.02 in individuals without overt hematologic disease, is currently considered a pre-cancerous state carrying a risk of converting to leukemia of 0.5 to 1% per year.[3,16]

Clonal hematopoiesis is thought to be rare in individuals younger than 50 years and increases in frequency with age, affecting at least 10% of individuals older than 65 years[2,4] and close to 20% of persons above 90 years.[5] Some mutations involved in clonal hematopoiesis, such as *JAK2* or *DNMT3A* mutations, have been detected in blood of healthy adults of all ages and are thought to be able to cause clonal expansion of mutant HSCs throughout life. On the other hand, a subset of recurrent mutations has only been observed in individuals over the age of 70 years, such as mutations in *SRSF2* or *SF3B1*, suggesting that clonal expansion of HSCs harboring these mutations is age-dependent.[5] This observation has led to the hypothesis that the aging cellular background may play a crucial role in the selection and expansion of mutant HSCs.[5] Indeed, aging is accompanied by a decline in HSC function,[17] a bias towards myeloid differentiation[18] and changes in the bone marrow niche.[19] It is therefore possible that certain mutations provide a cellular advantage in the aging bone marrow environment, allowing for clonal expansion of mutant HSCs exclusively in this context.[5,20,21] However, it is also possible that CHDMs associated with aging arise in young individuals but remain undetected due to technical limitations; because of the low VAF at which CHDMs are often present, the detection method used heavily influences the ability to identify these mutations.[22] For instance, studies favoring a targeted approach to provide deep sequencing coverage in known hotspot regions have led to the identification of a number of CHDMs with low VAF which would have otherwise been missed by exome or genome sequencing.[5,6]

In the present study, we aim to characterize the genetic profile and features of clonal hematopoiesis in individuals below the age of 70 years. To identify CHDMs with a VAF ≥0.002 in a cohort of over 2,000 population controls between 20 and 69 years of age, we use single molecule molecular inversion probes (smMIPs).[23–26] As a novel and highly flexible method for targeted enrichment of genomic regions of interest, we have made use of smMIPs to screen our cohort for somatic mutations in 40 established loci for CHDMs.

Furthermore, and unexpectedly, reference population databases for genetic variation have been found to contain pathogenic variants established to cause developmental disorders when present in the germline. A possible explanation for this surprising observation is that these mutations represent somatic mutations with elevated VAFs in blood due to their role as CHDMs.[27] However, the extent of the genetic overlap between somatic mutations in clonal hematopoiesis and germline mutations in developmental disorders remains unclear. Therefore, we screened our cohort for somatic mutations in an additional set of 64 loci in which recurrent germline *de novo* mutations have been found to cause severe developmental disorders. Several of these loci have been previously

implicated in paternal age effect disorders[28] with the causative mutations shown to cause clonal expansion in spermatogonial stem cells.[29–32] We here aim to determine whether these loci may represent novel sites for CHDMs.

Results

Sensitive and specific detection of somatic mutations in blood by smMIPs

We sequenced 104 loci in 2,007 blood samples of population controls between 20 and 69 years of age with a median unique coverage of 845-fold per sample (see Supplementary Figure S1 and S2). The median unique coverage corresponds to the number of unique DNA molecules sequenced per sample and per position after removal of PCR duplicates. Using an approach based on modeling sequencing error rates per targeted position, we identified 224 somatic SNVs and indels, of which 223 localize to the coding regions screened, with VAFs ranging between 0.0008 and 0.35 (average 0.015, median 0.0061). The median unique coverage of these mutation loci was 4103-fold.

To validate our findings, we used restriction digestion to analyze five somatic mutations localizing to a hotspot of *DNMT3A*, in which mutations disrupt a restriction digestion site for TauI (see Figure 1A and 1B). For all five samples, a band of undigested DNA was observed at 577 bp proportional in size to the mutation VAF per sample (Figure 1C and 1D). We selected one sample and one independent control for sequencing after restriction digestion and identified an enrichment for reads corresponding to the *DNMT3A* mutation compared to the wild-type allele for the sample in which a mutation was detected (Figure 1E). These findings support that mutations identified with VAFs as low as 0.002 represent true mutations in the original DNA sample rather than false-positives.

Somatic mutations in known clonal hematopoiesis driver genes in blood of population controls

In total, 216 somatic mutations were identified in coding regions of known clonal hematopoiesis driver genes in our cohort (see Figure 2A). Among these, 170 were mutations previously identified in clonal hematopoiesis.[2–5] For instance, several known clonal hematopoiesis driver missense mutations in genes such as *DNMT3A, JAK2*, *GNAS*, *NRAS*, *SRSF2* and *SF3B1*, were detected in our cohort (see Table 1). Additionally, nonsense mutations were identified in genes previously identified to harbor truncating mutations in clonal hematopoiesis such as *ASXL1*, *DNMT3A, TET2* and *TP53* and a total of 20 indels involving *DNMT3A* and *TET2* were detected (see Table 2). The most frequently mutated gene in our cohort is *DNMT3A,* for which a wide variety of mutations were observed including hotspot and non-hotspot missense mutations, loss-of-function point mutations

and indels (see Table 1, Table 2 and Figure 2B). Hotspots in *DNMT3A* are here defined as residues in which 5 or more missense mutations were detected in our cohort, including DNMT3A R326, R729, Y735, R736, W860 and R882.

Furthermore, somatic mutations were identified in known clonal hematopoiesis driver genes which had not been previously reported as CHDMs (see Table 3). These consist of 46 SNVs in coding regions of *BRAF*, *BRCC3*, *CBL*, *DNMT3A*, *GNAS, KRAS*, *NRAS, PIK3CA, PTPN11*, *TET2* and *TP53*. Several of these novel mutations were found adjacent to residues previously identified to harbor CHDMs. For instance, three individuals in our cohort were found to have KRAS G13D mutations. This missense mutation occurs contiguous to G12, a recurrently mutated residue in which G12C, G12R and G12S mutations have been identified in clonal hematopoiesis.[3,5] Additionally, we detected mutations within specific genes with a different mechanism than those usually observed to drive clonal hematopoiesis. For instance, while loss-of-function somatic *TET2* mutations are frequent in clonal hematopoiesis and leukemia, we detected one missense *TET2* mutation in a 20-year-old. Comparing the VAF of the 170 known and the 46 novel mutations we identified in genes involved in clonal hematopoiesis reveals a statistically significant difference in the VAF of both groups of mutations (0.0069 vs 0.0038 for known versus novel, p=0.0003, Wilcoxon rank sum test). This difference between both groups suggests that known mutations identified have a stronger clonal advantage than the novel mutations identified in our cohort, which may explain why they had not been detected previously.

Clonal hematopoiesis can arise throughout all adult life

The abovementioned 216 somatic mutations in clonal hematopoiesis-driver genes were identified in 192 individuals between 20 and 69 years of age, representing an overall prevalence of 9.56% for CHDMs throughout our cohort. Individuals in our cohort with clonal hematopoiesis were significantly older than those without (median 57 versus 43 years of age, $p<2.2x10^{-16}$, Wilcoxon rank sum test). The prevalence of CHDMs increased in an exponential manner with age and at least one CHDM was observed in 2.5% of the 20 to 29-year-old group, in 3.2% of the 30 to 39-year-old group, in 8.2% of the 40 to 49-year-old group, in 13.2% of the 50 to 59-year-old group and in 20.6% of the 60 to 69-year-old group (see Figure 3A). Two or more CHDMs were identified in 22 individuals in our cohort, who were significantly older than those in whom only one CHDM was identified (median of 65 versus 56 years of age, p <0.01, Wilcoxon rank sum test). Sixteen of these individuals with more than one CHDM were 60 or older, representing close to 4% of the group of individuals between 60 and 69 years of age (see Figure 3B). No difference in the mutational status or number of mutations was observed between sexes.

A

Lane	Age	Mutation gDNA	Mut/Coverage	VAF
1	62	chr2:25,457,242 C>T	15/7108	0.002
2	30	chr2:25,457,243 G>A	15/2449	0.006
3	42	chr2:25,457,242 C>T	55/6401	0.008
4	62	chr2:25,457,243 G>A	65/3164	0.021
5	61	chr2:25,457,242 C>T	146/2132	0.068

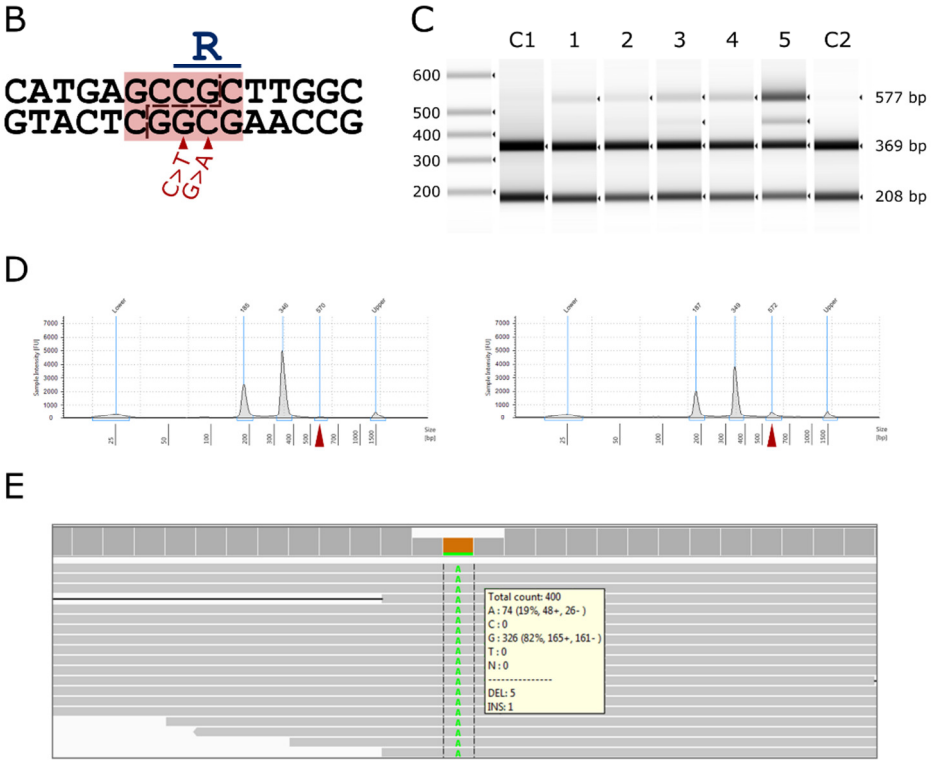


Figure 1. Validation of mutations by restriction digestion and re-sequencing. **A)** Mutations identified in DNMT3A R882 selected for validation. **B)** Scheme showing recognition site for restriction digestion enzyme TauI in the genomic sequence corresponding to DNMT3A R882. Mutations chr2:25457242C>T and chr2:25457243G>A leading to R882H and R882C respectively are marked below with red arrows. **C)** Size analysis of restriction digestion of *DNMT3A* PCR products. Lanes 1 to 5 represent samples with mutations with different VAFs. C1 is a control with false-positive signal for a G>A mutation at chr2:25457243, as determined statistically. C2 is a control with no *DNMT3A* mutation. **D)** Gel trace of size analysis of digestion, with sample 4 on the left and C2 on the right. The peak corresponding to the full size product is marked with a red triangle for both samples. Note that C2 present a small peak at 577 bp, corresponding to undigested PCR products due to digestion enzyme saturation. **E)** Sequencing results of digested PCR product for sample 5. We obtain a higher ratio of mutation to wild-type reads than in the original sample due to digestion of the wild-type product.

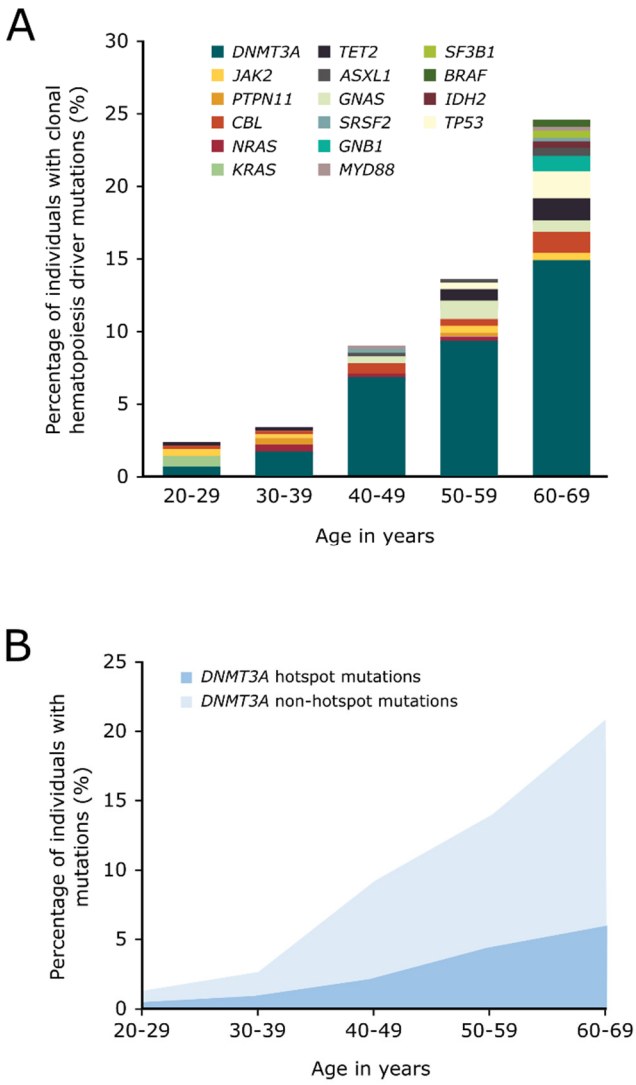


Figure 2. A) Prevalence and distribution of clonal hematopoiesis-driver mutations identified in healthy individuals aged between 20 and 69 years of age. **B)** Prevalence of mutations in *DNMT3A* per age group. Hotspot in *DNTM3A* are defined as residues in which 5 or more mutations were identified in our cohort and include R326, R729, Y735, R736, W860 and R882. All other missense, loss-of-function and indels are included in the non-hotspot mutations.

Remarkably, while the prevalence and number of CHDMs in the population increased with age, we did not detect a correlation between the VAF of CHDMs and the age of the individual in which they were identified (R^2=-0.03, Pearson's correlation; see Figure 3C). We examined a set of previously published CHDMs[5] in which we also observed a lack of correlation between the VAF of mutations identified in individuals below the age of 70 and the age of the

individual in which they were detected (R^2=0.06, Pearson's correlation). However, when analyzing the VAF of mutations in individuals 70 years of age or older, we observed a moderate correlation between these two parameters (R^2 = 0.3, Pearson's correlation). No statistically significant difference between the VAF of mutations identified in the abovementioned study in individuals below and above the age of 70 was observed (0.025 versus 0.027, p=0.87, Wilcoxon rank sum test).

Age-specific patterns of mutation in clonal hematopoiesis driver genes

Over 85% of all somatic SNVs in clonal hematopoiesis driver genes correspond to A>G/T>C, C>A/G>T or C>T/G>A mutations (170 out of 196 mutations). The distribution of these types of mutations differs between the individuals below the age of 45 and those 45 or older in our cohort (p<0.05, Pearson's Chi-squared test, df=2). While the most frequent CHMDs are C>T/G>A transitions, their frequency does not change with age, as they represent 51% and 52% of all SNVs in the younger and older age group. On the other hand, A>G/T>C mutations increase with age, corresponding to 20% of all SNVs in individuals 45 years or older in contrast with only 8% in individuals below this age (see Figure 4). This increase in frequency occurs at the expense of C>A/G>T mutations, which represent close to 30% of all SNVs in young individuals, as opposed to approximately 14% in individuals above the age of 45. This difference in patterns of mutations could not be attributed to mutations in any single gene.

Aging is associated with different patterns of clonal hematopoiesis and it has been proposed that *DNMT3A* and *JAK2* mutations appear throughout life, while mutations in *SRSF2* and *SF3B1* arise only in individuals over the age of 70.[5] Indeed, mutations in *DNMT3A* and *JAK2* were observed throughout all age groups; the youngest individual in which an established CHDM was identified in our cohort was a 22-year old woman with a *JAK2* V617F mutation at a VAF of 0.003. As with CHDMs in general, the frequency of *DNMT3A* mutations increased with age, reaching a prevalence of 13.7% in population controls between 60 and 69 years of age. The comparison of *DNMT3A* hotspot versus non-hotspot mutations reveals that the VAF of clones carrying *DNMT3A* hotspot mutations is significantly larger than that of clones with non-hotspot mutations (median 0.025 vs 0.009, Wilcoxon test, p<0.01). For most genes in which 5 or more CHDMs were identified in our cohort, including *CBL*, *DNMT3A*, *GNAS*, *JAK2* and *TET2*, no statistically significant difference was detected between the age of the carriers of mutations in any of these genes and individuals with one or more CHDMs in other genes. Although individuals with CHDMs in *TP53* were older than carriers of CHDMs in other genes (63.2 vs 53.9 years of age, Wilcoxon test, p <0.05), once we corrected for multiple testing this observation lost statistical significance at

Age	Gene	Mutation	VAF (%)
42	*ASXL1*	Q588*	2.23
55[e]	*ASXL1*	R404*	0.76
60	*ASXL1*	R693*	0.29
66[m]	*ASXL1*	R693*	0.91
24	*CBL*	R420Q	15.01
44	*CBL*	F418L	0.27
45[c]	*CBL*	C404Y	0.61
61	*CBL*	C404Y	1.29
63	*CBL*	G415V	1.15
25	*DNMT3A*	Y735S	0.64
25	*DNMT3A*	R326C	1.64
30	*DNMT3A*	C861*	0.50
30	*DNMT3A*	R882C	0.61
31	*DNMT3A*	R326C	0.61
35	*DNMT3A*	W860R	0.59
36	*DNMT3A*	S770L	1.62
36	*DNMT3A*	R736C	8.37
40	*DNMT3A*	Y735C	1.94
40	*DNMT3A*	Q886*	0.29
41[a]	*DNMT3A*	R882H	0.25
41[a]	*DNMT3A*	S770L	4.69
42	*DNMT3A*	R882H	0.86
42	*DNMT3A*	R887*	0.23
42	*DNMT3A*	M761V	0.38
44	*DNMT3A*	R771*	1.35
44	*DNMT3A*	R736C	2.09
44	*DNMT3A*	R882H	0.61
45[b]	*DNMT3A*	R771*	0.26
45	*DNMT3A*	R729W	1.79
45	*DNMT3A*	R771*	0.19
47	*DNMT3A*	Q886*	0.52
47	*DNMT3A*	R326C	1.74
47	*DNMT3A*	c.2322+3A>C	0.38
47	*DNMT3A*	S770*	2.04
49	*DNMT3A*	W860R	4.29
49	*DNMT3A*	Y735C	1.18
50	*DNMT3A*	E774*	0.42
51	*DNMT3A*	R882H	1.25
51	*DNMT3A*	R736C	0.25
51	*DNMT3A*	Y735C	5.72
51	*DNMT3A*	F732L	0.81
52	*DNMT3A*	R736C	1.88
53	*DNMT3A*	L737F	0.29

Age	Gene	Mutation	VAF (%)
54	*DNMT3A*	Y735C	0.46
54	*DNMT3A*	Y735C	0.58
55[d]	*DNMT3A*	L773R	0.92
55[d]	*DNMT3A*	R771G	0.20
55[e]	*DNMT3A*	G332E	0.16
55	*DNMT3A*	R729G	0.18
55	*DNMT3A*	R771*	0.28
56	*DNMT3A*	R736C	0.73
56	*DNMT3A*	W860R	9.78
56	*DNMT3A*	R882H	0.16
56	*DNMT3A*	R326H	0.43
57	*DNMT3A*	R736H	0.44
57	*DNMT3A*	W860R	1.94
57	*DNMT3A*	c.2597+2T>C	0.31
58	*DNMT3A*	R882C	0.18
59	*DNMT3A*	R882H	0.97
59	*DNMT3A*	R326C	0.63
59	*DNMT3A*	R882C	0.40
59	*DNMT3A*	R729W	0.28
60	*DNMT3A*	R736H	1.03
60	*DNMT3A*	R771*	0.92
60	*DNMT3A*	R326G	0.61
60	*DNMT3A*	R326C	11.08
60	*DNMT3A*	F734L	0.36
61	*DNMT3A*	W860*	0.61
61	*DNMT3A*	R882H	6.85
61	*DNMT3A*	R882H	6.06
61	*DNMT3A*	I769L	0.12
62	*DNMT3A*	W860R	0.09
62	*DNMT3A*	R882C	2.05
62	*DNMT3A*	R882H	0.21
62	*DNMT3A*	W330*	0.38
62	*DNMT3A*	S770L	2.79
62	*DNMT3A*	R326C	14.43
63	*DNMT3A*	R736C	0.75
63	*DNMT3A*	R736H	2.21
63	*DNMT3A*	R729Q	0.33
63	*DNMT3A*	Y735C	0.78
64	*DNMT3A*	R326S	35.02
64	*DNMT3A*	I769M	0.12
64	*DNMT3A*	R736H	0.33
65[j]	*DNMT3A*	R771*	0.93
65[k]	*DNMT3A*	R326H	1.02
65	*DNMT3A*	F772C	2.84
65	*DNMT3A*	R326H	1.25

Age	Gene	Mutation	VAF (%)
65	*DNMT3A*	R729W	0.46
66[m]	*DNMT3A*	R771*	0.34
66	*DNMT3A*	R320*	0.35
66	*DNMT3A*	R326H	0.78
66	*DNMT3A*	Q886*	2.58
66	*DNMT3A*	Y735C	2.09
67	*DNMT3A*	M880V	0.21
68[n]	*DNMT3A*	R771Q	2.15
68[p]	*DNMT3A*	R729G	0.18
68[q]	*DNMT3A*	Y735C	0.66
68	*DNMT3A*	E774K	1.24
68	*DNMT3A*	M761V	0.69
68	*DNMT3A*	c.2597+1G>A	2.16
69[r]	*DNMT3A*	I769S	0.12
69[r]	*DNMT3A*	M880I	0.53
69[r]	*DNMT3A*	R736H	0.46
69	*DNMT3A*	Y735C	1.91
45[c]	*GNAS*	R844H	2.10
50	*GNAS*	R844C	0.37
56	*GNAS*	R844H	1.59
56	*GNAS*	R844H	0.32
57	*GNAS*	R844S	0.55
58	*GNAS*	R844C	0.65
62	*GNAS*	R844H	2.32
63[h]	*GNAS*	R844H	0.52
60	*GNB1*	K57E	0.76
63[h]	*GNB1*	K57E	0.41
65	*GNB1*	K57E	0.73
69	*GNB1*	K57E	2.17
62	*IDH2*	R140Q	0.58
62	*IDH2*	R140Q	1.16
22	*JAK2*	V617F	0.31

Age	Gene	Mutation	VAF (%)
26	*JAK2*	V617F	0.93
30	*JAK2*	V617F	4.65
54	*JAK2*	V617F	9.73
55	*JAK2*	V617F	5.71
64[i]	*JAK2*	V617F	2.72
65[j]	*JAK2*	V617F	0.64
45[b]	*MYD88*	L273P	0.20
68[o]	*MYD88*	L273P	0.91
47	*NRAS*	Q61K	0.33
56	*NRAS*	Q61L	0.10
39	*PTPN11*	E69A	0.19
60[f]	*SF3B1*	K666N	0.67
68[n]	*SF3B1*	K700E	1.01
49	*SRSF2*	P95L	9.01
66	*SRSF2*	P95H	1.30
38	*TET2*	R550*	0.64
51	*TET2*	R550*	0.69
54	*TET2*	R550*	0.37
56	*TET2*	R550*	0.72
60[g]	*TET2*	Q764*	0.13
62	*TET2*	Q530*	0.81
64	*TET2*	R550*	0.51
67	*TET2*	Q548*	0.44
67	*TET2*	Q886*	1.58
53	*TP53*	R110L	0.17
62	*TP53*	R213*	0.29
63	*TP53*	L114*	0.34
67	*TP53*	Y220C	1.04
60[f]	*U2AF1*	S34F	0.98

Table 1. Known clonal hematopoiesis driver SNVs identified. Samples with more than one mutation are marked with a superscript letter for identification.

Age	Mutation (hg19 gDNA)	Gene	mRNA change	Protein change	VAF (%)
33	chr2:25458597del	*DNMT3A*	c.2576del	L859Yfs	1.67
42	chr2:25463300dup	*DNMT3A*	c.2193dup	F732Lfs	4.05
48	g.25458620_25458627dup	*DNMT3A*	c.2546_2553dup	M852Lfs	4.76
49	chr2:25463297dup	*DNMT3A*	c.2196dup	E733*	2.56
49	chr2:25463196_25463197del	*DNMT3A*	c.2297_2296del	K766Efs	0.32
57	chr2:25458596dup	*DNMT3A*	c.2577dup	W860Mfs	0.45
57	chr2:25458591delinsTT	*DNMT3A*	c.2582delinsAA	C861*	0.70
58[b]	chr2:25463196_25463197del	*DNMT3A*	c.2297_2296del	K766Efs	13.60
60	chr2:25463308del	*DNMT3A*	c.2185del	R729Gfs	5.86
61	chr2:25463196_25463197del	*DNMT3A*	c.2297_2296del	K766Efs	7.17
63	chr2:25463312del	*DNMT3A*	c.2181del	G728Afs	2.05
64	chr2:25463196_25463197del	*DNMT3A*	c.2297_2296del	K766Efs	0.33
65[k]	chr2:25463297dup	*DNMT3A*	c.2196dup	E733*	0.63
66[s]	chr2:25463291del	*DNMT3A*	c.2202del	Y735Tfs	0.90
66[s]	chr2:25458605_25458606del	*DNMT3A*	c.2568_2567del	E856Gfs	0.30
67	chr2:25458608del	*DNMT3A*	c.2565del	E856Rfs	0.91
68	chr2:25463196_25463197del	*DNMT3A*	c.2297_2296del	K766Efs	0.76
68[t]	chr2:25463190del	*DNMT3A*	c.2303del	D768Afs	0.16
68	chr2:25458595del	*DNMT3A*	c.2578del	W860Gfs	0.12
64	chr4:106157389dup	*TET2*	c.2290dup	Q764Pfs	0.44

Table 2. Indels identified in known clonal hematopoiesis driver genes. Samples with more than one mutation are marked with a superscript letter for identification.

▶ **Table 3. Novel somatic mutations identified in coding regions of clonal hematopoiesis driver genes.** Samples with more than one mutation are marked with a superscript letter for identification.

Age	Gene	Mutation	VAF (%)	Mutation type	PhyloP	CADD PHRED
67	*BRAF*	F595L	0.09	Missense	2.26	24.7
67	*BRAF*	K601N	0.98	Missense	1.12	21.3
54	*BRCC3*	D88G	2.94	Missense	8.54	17.8
35	*CBL*	P395H	0.17	Missense	7.45	19.7
41	*CBL*	C384Y	1.58	Missense	9.42	18.4
52	*CBL*	H398Q	0.81	Missense	0.95	14.8
58	*CBL*	C419S	0.32	Missense	9.48	21.4
69	*CBL*	S376P	0.97	Missense	7.66	18.2
65[v]	*CBL*	H398Q	0.27	Missense	0.95	14.8
65[v]	*CBL*	T377I	1.75	Missense	7.45	18.1
68[t]	*CBL*	C396Y	0.15	Missense	9.42	18.9
41[a]	*DNMT3A*	V328F	0.53	Missense	7.57	27.8
48	*DNMT3A*	D768E	0.33	Missense	5.68	24.7
49	*DNMT3A*	L888P	0.43	Missense	9.34	18.7
49	*DNMT3A*	M864R	0.21	Missense	9.23	24.9
49	*DNMT3A*	T862I	0.36	Missense	9.75	32
51	*DNMT3A*	D765G	0.90	Missense	8.04	24.9
51	*DNMT3A*	V763G	0.11	Missense	7.33	25.5
52	*DNMT3A*	A884T	0.28	Missense	6.09	28.7
54	*DNMT3A*	E774E	0.42	Synonymous	7.78	15.6
54	*DNMT3A*	R885S	0.28	Missense	1.59	18.1
55	*DNMT3A*	K766N	1.10	Missense	2.56	19.6
57	*DNMT3A*	A884V	1.27	Missense	9.86	28
59	*DNMT3A*	R866M	0.16	Missense	7.73	35
61	*DNMT3A*	T862I	0.52	Missense	9.75	32
64	*DNMT3A*	D857A	1.32	Missense	7.95	28.4
68	*DNMT3A*	D768E	0.38	Missense	5.68	24.7
68	*DNMT3A*	L889V	0.35	Missense	7.99	19.8
64[j]	*DNMT3A*	E863G	0.26	Missense	7.95	32
68[q]	*DNMT3A*	T862I	0.80	Missense	9.75	32
46	*GNAS*	C843R	0.46	Missense	7.55	21.2
62	*GNAS*	L846P	0.42	Missense	7.55	21.2
22	*KRAS*	G13D	1.07	Missense	7.74	27.8
25	*KRAS*	G13D	0.37	Missense	7.74	27.8
28	*KRAS*	G13D	0.39	Missense	7.74	27.8
35	*NRAS*	E62*	0.08	Nonsense	7.55	38
37[u]	*NRAS*	E62*	0.12	Nonsense	7.55	38
23	*PIK3CA*	E39*	0.08	Nonsense	9.41	29
37[u]	*PTPN11*	L65L	0.16	Synonymous	0.62	9.4
56	*PTPN11*	T73A	0.11	Missense	9.33	19.5
20	*TET2*	Q548K	1.68	Missense	0.57	4.4
60[g]	*TP53*	F113V	0.08	Missense	7.39	24.6
63	*TP53*	M237I	0.52	Missense	7.56	22.8
69	*TP53*	V216M	0.39	Missense	7.78	28.9
68[o]	*TP53*	M237K	0.31	Missense	9.02	24
68[p]	*TP53*	V216M	0.90	Missense	7.78	28.9

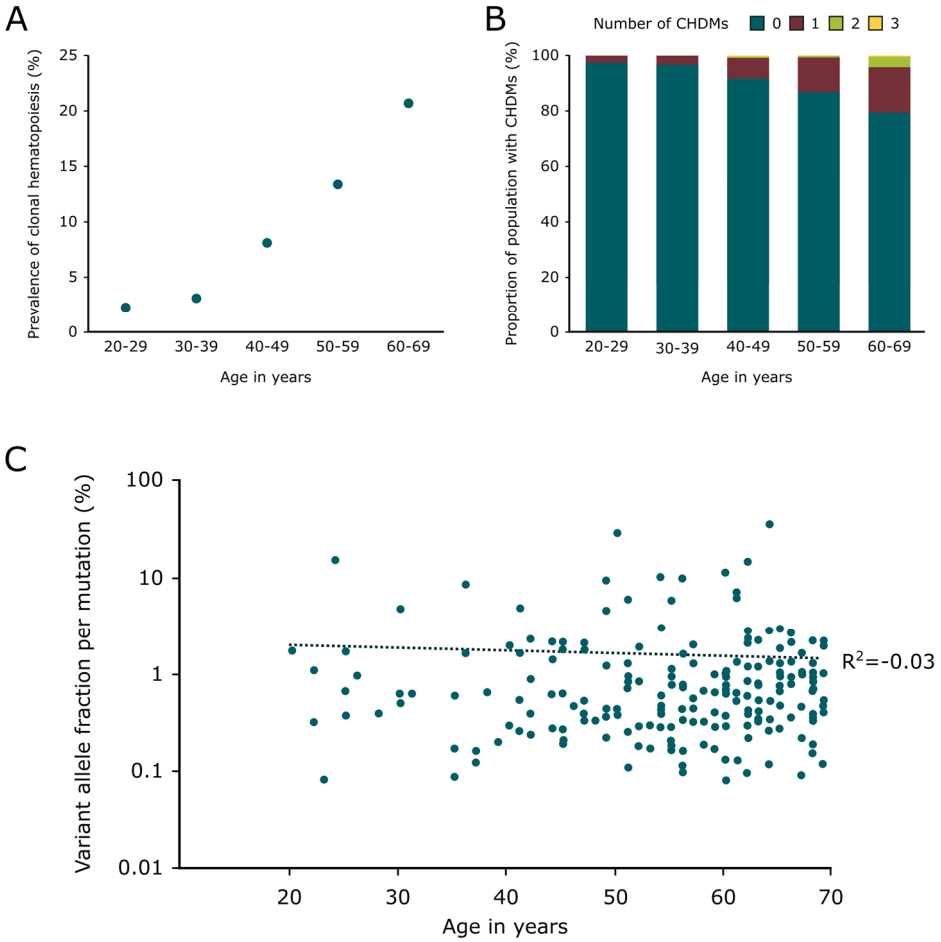


Figure 3. A) Prevalence of clonal hematopoiesis per age group, defined as the frequency of individuals with one or more CHDMs per decade of age. . B) Proportion of the population per age group with CHDMs.. C) Variant allele fraction per mutation identified per age of the individual in which the mutation was identified. The y-axis is in logarithmic scale. No correlation is observed between the age of the individual and the VAF of the mutation identified. A large number of low VAF mutations are detected in older individuals.

$p<0.05$. Some genes included in our screen were found to be mutated only in individuals over the age of 60, such as *GNB1*, *IDH2*, *SF3B1* and *U2AF1*. Although mutations in components of the spliceosome have been proposed to drive clonal hematopoiesis only in older individuals, *SRSF2* mutations were identified in two individuals of 49 and 66 years of age. Remarkably, the mutation observed in the 49-year old man was a P95L missense mutation in *SRSF2* with a relatively high VAF of 0.09.

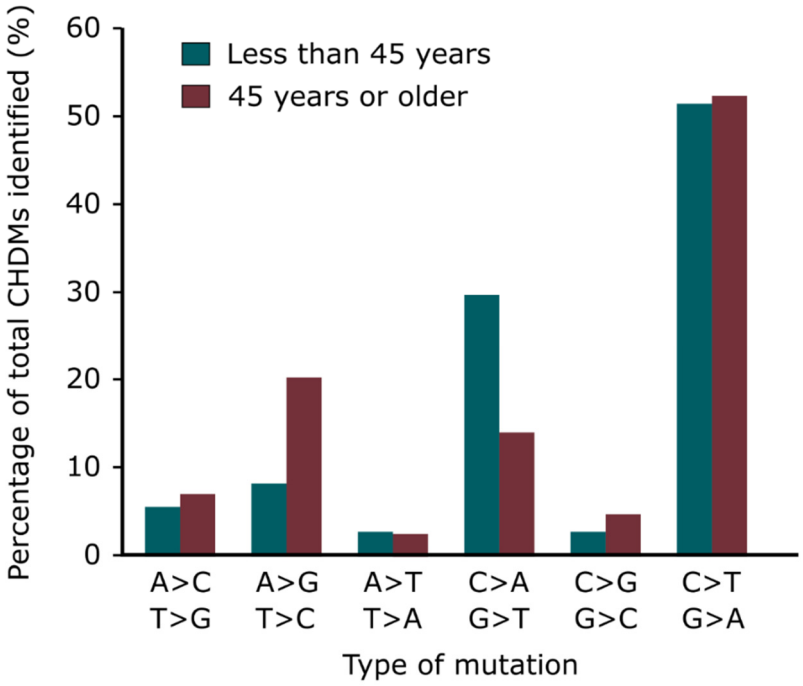


Figure 4. Type of mutation change per age group

Somatic mutations in candidate loci for clonal hematopoiesis driver mutations

Unexpectedly, mutations known to cause developmental disorders when arising in the germline can be found at low frequencies in population databases of genetic variation such as ExAC.[27] It has been suggested that some of these findings may be explained by mutations which, in addition to causing developmental disorders when present in the germline, would also be involved in clonal hematopoiesis when arising in HSCs.[27] A revision of published literature led to the identification of 1,158 somatic mutations in 90 genes involved in clonal hematopoiesis.[2–5] We compared this list of 90 genes with a subset of 464 genes involved in dominant and *de novo* developmental disorders from DDG2P[33] to determine whether there was significant overlap between genes involved in clonal hematopoiesis when mutated somatically and genes involved in developmental disorders. In this way, we identified 29 genes in which germline mutations cause developmental disorders while somatic mutations are involved in clonal hematopoiesis. This represents a significant enrichment compared to the expectation (hypergeometric distribution with a universe of 22,285 protein-coding genes, $p = 9.2 \times 10^{-28}$).

Age	Gene	Mutation	VAF (%)	Mutation type	PhyloP	CADD PHRED	Grantham Score
25	*COL4A3BP*	S260*	0.12	Nonsense	9.84	40	NA
41	*ADNP*	Y719Y	0.25	Synonymous	-1.96	0.02	NA
43	*KCNQ3*	S228N	0.30	Missense	5.91	19.7	46
48	*RHEB*	Y35Y	1.73	Synonymous	0.10	8.8	NA
58	*CUX2*	L592P	2.46	Missense	8.04	21.2	98
59	*HECTD1*	S2223P	0.12	Missense	8.96	12.6	74
60	*SMAD4*	C499C	0.17	Synonymous	1.86	4.9	NA

Table 4. Somatic mutations identified in candidate clonal hematopoiesis driver genes.

In addition to the 40 known loci for CHDMs in our assay, we screened for somatic mutations in blood in 64 novel candidate loci for CHDMs. These included seven loci with nine residues in which recurrent missense mutations are known to be causative for paternal age effect disorders, such as achondroplasia, when present in the germline.[29–31,34,35] These mutations cause clonal expansion in spermatogonial stem cells when arising during spermatogenesis[28] and we hypothesized that some may also undergo clonal expansion in other tissues. Therefore, we considered mutations in these loci as candidates for clonal expansion in HSCs during hematopoiesis. Furthermore, we selected 57 additional loci in which recurrent identical *de novo* mutations have been found in developmental disorders.[36–38] Screening of these 64 candidate loci for CHDMs led to the identification of 7 somatic mutations, including 1 nonsense, 3 missense and 3 synonymous mutations with VAFs ranging from 0.0012 to 0.024 (average 0.0074, median 0.0025). Genes in which mutations were detected include *ADNP*, *COL4A3BP*, *CUX2*, *HECTD1*, *KCQN3, RHEB* and *SMAD4* (see Table 4). No mutations were identified that were identical to mutations involved in spermatogonial stem cell selection or found to be recurrently mutated in developmental disorders.

Discussion

Recurrent somatic mutations have been recently identified in blood-derived DNA of population controls.[2–4] In this study, we established a smMIPs assay targeting 104 known or candidate loci for CHDMs and sequenced blood-derived DNA of 2,007 population controls between the age of 20 and 69. With a median unique sequencing coverage of 845-fold, corresponding to the number of unique DNA molecules captured per position per sample, we identified 223 somatic SNVs and indels in the coding regions screened of which 216 affected known clonal hematopoiesis-driver genes. A high ratio for non-synonymous to synonymous somatic mutations was observed for loci screened within genes known to drive clonal hematopoiesis (214 non-synonymous versus 2

synonymous point mutations). This observation suggests overall positive selection among a large proportion of the mutations identified in the screened loci.[39]

Both the prevalence of clonal hematopoiesis as well as the number of mutations increased with age, reaching a frequency over 20% for CHDMs in individuals between 60 and 69 years of age. This number is close to double that of previous reports mentioning a prevalence for CHDMs in 5 to 10% in healthy individuals older than 60.[2–4] It is likely that the higher prevalence of clonal hematopoiesis detected in our study results from the increased sensitivity of smMIPs, compared to other sequencing methods used in previous studies.[22] The use of a deep coverage smMIP approach has the advantage of allowing a large number of loci to be screened for mutations with high sensitivity for the identification of mutations with low VAFs. In contrast, exome sequencing has the advantage of allowing the identification of mutations throughout the coding region, but it may miss mutations present at low allelic frequencies when performed at current day standard coverage. Indeed, an average sequence of 84 reads sets the lower limit for detection of somatic mutations at an allele fraction of approximately 0.035.[3] More than 90% of the CHDMs detected in our study (198 out of 216) are below this VAF and would most likely have been missed by average coverage exome sequencing. On the other hand, amplicon-based sequencing limits the detection of mutations to the targeted regions but provides deep coverage, which enables the identification of mutations with low VAF. For instance, one study detected mutations in 15 loci using a sequencing coverage above 1000-fold, which lowered the limit for mutation detection to an allele fraction of 0.008.[5] One of the main limitations for the detection of somatic mutations is the sequencing error rate of NGS platforms, ranging around 0.1-1% for sequencing by synthesis approaches.[40] Mutations introduced during the preparation of the sequencing library further limit the ability to distinguish true CHDMs present in blood from false-positive signal.[41] Some of these limitations can be bypassed by using smMIPs, which include a random tag in each probe assigning a unique molecule identifier (UMI) to each individual DNA molecule captured.[23] Through this, multiple sequencing reads descending from the same DNA molecule can be traced in order to generate a true molecular count without PCR duplicates. Additionally, this can be used to create a consensus sequence of this DNA fragment, thus increasing the specificity for the detection of true somatic events. To ensure specificity in our detection of CHDMs, we modeled the sequencing error to identify statistically significant mutation counts in two replicates from the same sample. Furthermore, at least two independent smMIPs were used to target each DNA strand within the screened loci. The mutations identified in our assay had a median VAF of 0.0061, ranging from 0.0008 to 0.35, and a subset of mutations were validated using restriction digestion, confirming that the deviations in number of mutation reads at specific positions detected through sequencing data reflected a true mutation present in the sample DNA.

Although the prevalence of clonal hematopoiesis increased with age, CHDMs were observed throughout all ages in our cohort. The youngest individual in which an established CHDM was identified in our cohort was a 22-year-old woman in whom we detected a *JAK2*V617F mutation with a VAF of 0.003. CHDMs were detected in 2.3% of population controls between 20 and 29 years of age, suggesting that clonal hematopoiesis is not a rare finding in early adult life. While age was a poor predictor for the presence and number of CHDMs at the individual level, when analyzing our findings by ten-year age groups, we observed that age explained 96% of the variation in the presence and number of CHDMs in the population. This suggests that while age is a major factor in the occurrence of CHDMs at the population level, stochasticity and inter-individual differences, such as genetic variation or environmental factors, may play a prominent role in the occurrence of CHDMs at the individual level. Notably, we observe that the increase in prevalence and number of CHDMs at the population level is not linear. This supports that factors associated with aging may accelerate the occurrence of CHDMs over time, either due to increased DNA damage or to decreased DNA repair. An increase with age was observed in the proportion of A>G:T>C mutations detected in our cohort. This type of mutation has been linked to deamination of adenine and has been shown to be associated with transcription-coupled repair.[42] The prevalence of CHDMs in population controls contrasts with the incidence of leukemia in the general population. Given this discrepancy, it is unclear whether all individuals with CHDMs have an increased risk for developing leukemia. The presence of CHDMs may place individuals at increased risk for development of hematologic malignancies, but does not seem sufficient in itself to develop leukemia. It is therefore likely that the evolution from clonal hematopoiesis to leukemia is the result of many cycles of selection of mutant HSCs in which the cellular, tissue and organism environment plays a role.

While some CHDMs, such as those in *DNMT3A* and *JAK2*, were found in individuals of all ages, some CHDMs were exclusively detected in individuals over the age of 60, including mutations in *GNB1*, *IDH2*, *SF3B1* and *U2AF1*. Additionally, individuals with CHDMs in *TP53* were older than those with CHDMs in other genes, although this result did not reach statistical significance. It has been suggested that some mutations may arise in HSCs in young individuals but can only expand clonally in the context of the aging bone marrow.[5,13,20] Mutations in components of the spliceosome, such as *SRSF2* and *SF3B1,* have been proposed to fall under this category.[5] Interestingly, we identified a *SRSF2* mutation with a VAF of 0.09 in a 49-year-old individual, suggesting that although rare, clonal expansion of spliceosome mutations can occur in younger individuals. Compared to HSCs in young individuals, aging HSCs show decline in replication capacity[17,22] and a bias towards myeloid differentiation.[18] These changes may be specific to the aging cells or could result from alterations in the bone marrow niche[17,19,43]. Cell-intrinsic alterations arising in aging HSCs include epigenetic alterations and upregulation of genes involved in myeloid differentiation, DNA repair, cell death and genes linked to leukemia.[14,17,43] However, computational modeling and

studies in mouse models support that the aging microenvironment exerts strong selective pressure on HSCs.[11,22] Therefore, certain mutations involved in clonal hematopoiesis may confer aging HSCs a growth advantage over wild-type aging HSCs in which function is generally declining. In contrast, the same mutations in young HSCs may not result in increased cellular fitness compared to wild-type young HSCs and thus hamper expansion.[44]

The gene harboring most CHDMs in our study was *DNMT3A,* where we identified hotspot and non-hotspot missense mutations, as well as truncating mutations throughout all five loci screened. Our assay covered approximately 8.5% of the coding sequence of *DNMT3A* and mutations in this gene have been identified throughout its coding sequence. Similarly, our assay only included close to 2.5% of the coding sequence of *ASXL1* and *TET2*, which are disrupted by truncating mutations that may arise throughout the gene. Therefore, the detected prevalence of somatic mutations in *DNMT3A, ASXL1* and *TET2* in our study is likely an underestimation and more mutations may be found outside the regions screened. Nevertheless, our findings support *DNMT3A* as the most frequently mutated gene in clonal hematopoiesis.[2–5] Close to 60% of *DNMT3A* mutations identified in hematologic malignancies disrupt codon R882,[45,46] while mutations involving codon R326 represent only 0.1% of *DNMT3A* mutations in hematologic cancer.[47] In clonal hematopoiesis, we find that 9.6% of *DNMT3A* mutations identified in controls disrupt codon R882, similar to those disrupting R326 which represent 8.9% (13/135 and 12/135 mutations for R882 and R326, respectively). We therefore estimate that the ratio of DNMT3A R882 to R326 mutations is approximately 600:1 in malignancy and observe that this ratio is 1.1:1 in clonal hematopoiesis. This supports the hypothesis that recurrent *DNMT3A* mutations grant an advantage to mutated cells, but the risk of progressing to malignancy is much higher in the presence of DNMT3A R882 mutations.[45] This higher risk associated with these mutations may stem from the dominant negative effect of DNMT3A R882 mutants which severely affects DNMT3A function through homodimeric interactions with the wild-type protein.[48]. Mutations in *DNMT3A* affecting residues other than R882 are thought to have a smaller effect on the activity of the wild-type protein, which may not be sufficient to drive the development of cancer.[48] Residues other than R882 can also be mutated in myeloid and lymphoid malignancies and the bi-allelic presence of this type of *DNMT3A* mutations has been identified in different forms of leukemia. In our cohort, we identified six individuals with more than one *DNMT3A* mutation but the VAF reflected the existence of clones of different sizes. It is unclear whether in these subjects the two mutations may be present in the same cell or whether they represent completely independent *DNMT3A* mutant clones.

No significant correlation was observed between the VAF of the CHDMs identified and the age of the individuals carrying the mutations (R^2=-0.03, Pearson's correlation). This may result from an increase in the occurrence of mutations over time, leading to a higher number of mutant clones of small size in

older individuals that would lower the average VAF in this group. However, the fact that examining only mutations with a VAF ≥ 0.02 also reveals a lack of correlation between age and VAF argues against this point. Another explanation lies in the possibility that the size of a mutant clone depends on the age of the clone, rather than the age of the individual in which it is present. Our study consisted of a single measurement and as such, we are not able to follow the evolution of mutant clones over time. One paper analyzing multiple samples from the same individual followed the evolution of mutant clones over time and determined that all mutant clones were still present between 4 and 8 years after initial detection, with the vast majority of mutant clones increasing or remaining stable over time.[3] These fluctuations may reflect actual changes in the size of mutant HSCs clones, but could also be due to variation over time in the contribution of mutant clones of HSCs to blood. It is unclear at present whether the size of a mutant clone as reflected by the VAF of a mutation does in fact increase with the age of the clone. The presence of a CHDM in itself may not be sufficient to lead to clonal expansion over time, either because of a weak proliferative effect or because an additional factor may be required to allow for expansion. A higher correlation between VAF and age was observed for individuals over the age of 70 from a different study [5], which may suggest that aging HSCs and bone marrow environment may represent this additional factor that allows for clonal expansion of mutant HSCs.

A recent study highlighted the unexpected presence of mutations associated with developmental disorders at high frequency in 60,706 reference exomes in ExAC.[27] For instance, 345 individuals were found to carry 56 different truncating *ASXL1* mutations, an unexpected finding, considering that germline truncating *ASXL1* mutations cause Bohring-Opitz syndrome, a severe developmental syndrome.[27,49] This study shows that these mutations had lower VAFs than expected for germline events and were present mainly in older individuals and therefore likely represented CHDMs. This suggests that the presence in ExAC of other mutations causative for developmental disorders may also reflect somatic events involved in clonal hematopoiesis rather than germline mutations.[27] We therefore screened our cohort for somatic mutations in candidate genes for clonal hematopoiesis which cause developmental disorders when mutated in the germline. Similar to this previously published study, we identified somatic mutations in blood overlapping with germline mutations known to cause developmental disorders when present in the germline. In addition to four truncating mutations in *ASXL1*, we identified 13 missense mutations in DNMT3A R882, which is known to be mutated in Tatton-Brown syndrome, a developmental disorder with overgrowth.[50] Similarly, 3 somatic missense mutations in *CBL* overlapping with germline mutations leading to a Noonan-like phenotype were identified in our cohort.[51,52] However, we failed to identify somatic mutations in blood in our candidate loci which overlapped with known developmental disease-causing mutations. This suggests that this genetic overlap between developmental disorders and clonal hematopoiesis may be

restricted to specific mutations in known candidate genes for clonal hematopoiesis. Thus, the presence of other developmental disease-causing variants in reference databases remains unexplained and may also be due to other factors such as sequencing errors. We did, however, identify a somatic mutation in *CBL* leading to CBL R420Q at a VAF of 0.15 in a 24-year-old man, entailing that 30% of circulating blood cells carry this mutation. The high VAF for this mutation is remarkable given the young age of the individual and the fact that this was the only somatic mutation identified in this individual. As such, it is unknown whether this mutation represents a somatic event arising in a HSC during postnatal life or a postzygotic *de novo* mutation arising in early embryogenesis. Interestingly, this mutation has been reported to cause Noonan-like syndrome when present in the germline[51] and leukemia when present somatically.[47] We could not access clinical information or additional samples from this individual to verify the presence of mosaicism in other tissues. It may be that pathogenic mutations such as the one identified in this individual or those present in these reference databases reflect genetic variation in resilient individuals.[53]

In summary, we have screened a cohort of population controls between 20 and 69 years of age to identify somatic mutations in blood implicated in clonal hematopoiesis. Our method provides high sensitivity which allowed for the identification of CHDMs at higher prevalence than previously reported despite studying a limited set of mutations. Somatic mutations were identified in individuals of all ages, with differences in the profile of genes mutated per age group and a strong increase in the number of mutations detected with age, with over 20% of individuals between 60 and 69 years carrying at least one CHDM, while up to 3% of individuals younger than 30 years of age had at least one CHDM. Our findings support the occurrence of clonal hematopoiesis associated with somatic mutations as a widespread mechanism linked to aging, suggesting that clonal evolution of cells harboring somatic mutations is a universal mechanism occurring at all ages in humans.

Methods and materials

Samples

This study was performed using data and biomaterial from the Nijmegen Biomedical Study (NBS). The NBS is a population-based study among 9,350 individuals, based on an age- and sex stratified random sample from the register of the municipality of Nijmegen, a city in the east of the Netherlands. Extensive questionnaire data on health and lifestyle were collected. Blood samples were collected in EDTA tubes and DNA was extracted by salt precipitation method.[54] For this study, we obtained DNA samples and information on age and sex for 2,014 NBS participants via the Radboud Biobank.[54] Approximately four hundred samples equally distributed between men and women were obtained for each age group (400 for age group 20-29, 405 for age group 30-39, 404 for age group 40-49, 403 for age group 50-59 and 402 for age group 60-69 years of age; see Supplementary Table S1). The quality of purified DNA was tested and each sample was normalized to 25 ng/µl by optical density measurement (Dropsense, Trinean). This study was approved by the Committee on Research involving Human Subjects (*Commissie Mensgebonden Onderzoek*) of the Radboudumc (CMO approval: 2015-2228).

Targeted loci to screen for mutations

We performed a literature review to identify mutations observed recurrently in age-related clonal hematopoiesis. By combining the results from several published studies,[2–5] we collected 1,158 coding mutations in 513 amino acid residues which were ranked by total number of mutations identified per residue. We selected 35 loci with the largest number of coding mutations observed in clonal hematopoiesis, which corresponds to a total of 599 single nucleotide variants (SNVs) in 87 residues. Furthermore, five loci in which CHDMs have been previously identified and in which overlapping germline or postzygotic *de novo* mutations are known to cause developmental disorders were included.[55–58] In addition, we selected seven loci with nine residues in which mutations are known to be causative for paternal age effect disorders and to lead to spermatogonial stem cell expansion.[28] Finally, we included 57 additional loci in which recurrent identical *de novo* mutations have been found in developmental disorders.[36–38] These loci are therefore candidates either for elevated mutation rates at that genomic site or for mutations leading to expansion of spermatogonial stem cells. This analysis resulted in total in 104 loci.

smMIP design

To screen for mutations in these 104 loci, MIPGEN software[59] was used to design probes covering the regions of interest, followed by manual curation and selection. The smMIPs were 80 nucleotide-long DNA molecules consisting of an

extension and ligation arm with a combined length of 40 nucleotides, separated by a linker sequence of 30 nucleotides (see Supplementary Figure S1 for more details). All smMIPs contained a unique molecule identifier (UMI) or molecular tag consisting of 2×5 random nucleotides to identify each individual captured DNA molecule. These smMIPs were designed to target regions of 54 nucleotides. At least one smMIP on the sense and on the antisense DNA strand were designed per locus. Probes targeting genomic regions containing a SNP with a population frequency above 1% were designed to have complementary arms targeting both alleles. In total, 231 smMIPs were designed to capture the 104 regions of interest. The smMIP oligonucleotides were produced by Integrated DNA Technologies (IDT, Leuven, Belgium) at 25nmol scale and normalized to a concentration of 100μM.

smMIPs assay setup

The smMIPs assay was set up with minor modifications to previously published protocols.[23,60] Briefly, individual smMIPs were pooled equimolarly and phosphorylated using T4 polynucleotide kinase and 10x T4 DNA ligase buffer supplemented with 10mM ATP (New England Biolabs). The smMIP capture was performed on 8μl of input DNA (200 ng) supplied with 17μl of capture mix containing 0.28μl of a phosphorylated smMIP pool dilution at 3.12nM, resulting in a ratio of 8,000 smMIP molecules per DNA molecule. The capture reaction was incubated for 18-22 hours at 60°C, after which the mix was cooled and treated with exonuclease. Each exo-treated sample was split in two technical replicates of 10μl, which were then amplified and barcoded separately by PCR. The PCR products were run on gel, pooled, purified using AMPureXP Beads (Agencourt) and run on a Tapestation (Agilent) to verify the integrity of the sequencing library. Sequencing was performed on an Illumina NextSeq500 platform with 2×79-bp paired-end reads (*i.e.* each DNA insert being sequenced in both directions). A pilot smMIP experiment was performed on control DNA for the optimization of the smMIP library. The performance of each individual probe was evaluated by examining the sequencing coverage per probe, in order to identify under-performing and over-performing smMIPs. After excluding smMIPs with off-target capture, a new and rebalanced smMIP pool was prepared adjusting the volumes for each probe. The new pool was phosphorylated and an experiment was run on control DNA samples to verify pool rebalancing. The library was subsequently prepared and sequenced as described previously using blood DNA samples from the cohort. After sequencing, 18 smMIPs were shown to have an overall median coverage <20-fold and were excluded from further analysis. Seven samples for which both replicates had an average sequencing coverage below 100-fold were excluded due to poor quality or quantity of the input DNA.

Analysis

FASTQ files were obtained from the bcl files and demultiplexed using the sample barcode. Sequenced reads were mapped with BWA MEM, using a modified version of an in-house bioinformatics pipeline which allows trimming of the MIP extension and ligation arms (MIPVAR, available at www.sourceforge.net/projects/mipvar/). To analyze unique DNA molecules, we detected and removed PCR duplicates of individual captured molecules per sample and per smMIP by identifying reads with the same UMI. From the regions selected for screening, 88 out of 104 loci had one or two mutated residues. For these loci, we analyzed the genomic region corresponding to the mutated residue(s) +/- 6 base pairs. Regions with more than two mutated residues (13 out of 104 loci) were analyzed so that we would examine the entire region encompassing all mutated residues within the locus +/- 6 bp. For *TP53*, we analyzed the entire region captured by the smMIPs (3 out of 104 loci). Pileups were generated for all samples for these positions using SAMtools with the following filter settings: sequence quality ≥25 and a mapping quality ≥15.[5] The coverage, each nucleotide change and the presence of insertions and deletions were counted at each position for each sample replicate. Within each sample replicate, positions with a unique molecule sequencing coverage below 200-fold were excluded. The sequencing error was calculated for each position and nucleotide change (A, C, G, T, insertion and deletion) based on all samples from the cohort. Additionally, the run-specific sequencing error for each position and nucleotide change was determined using only samples within the same sequencing run. Subsequently, we used a Poisson distribution to calculate a p-value reflecting the probability to obtain a number equal or higher to the observed number of mutation reads per position for all sample replicates based on the sequencing error. This calculation was performed in parallel using each determined sequencing error in two independent analyses. For each sample, we extracted the p-values for all nucleotide changes for all positions from only one of the replicates and performed Benjamini-Hochberg multiple test correction on these values. All nucleotide changes with an adjusted p-value <0.05 and with ≥2 reads for SNVs and ≥5 reads for indels were included as "statistically significant nucleotide changes". We then extracted from the second sample replicate the p-values obtained for the statistically significant nucleotide changes in the first replicate and performed Benjamini-Hochberg multiple test correction. We used the same filtering criteria and obtained a list of potential mutations. Finally, the list of potential mutations obtained with the overall and the run-specific sequencing error were overlapped. We included only mutations in which both replicates have ≥2 reads for SNVs and ≥5 reads for indels, representing a statistically significantly higher number of mutations counts than expected based both on the overall and the run-specific sequencing error. For positions within a mutational hotspot in which more than 5 mutations were identified (such as *JAK2* V617F and all *DNMT3A* hotspots), a separate analysis examining only nucleotide changes at those hotspots was performed. To exclude germline events, somatic

mutations were defined as mutations with a VAF ≤0.35 and an allele frequency <0.001 in ExAC.

Validations

Mutations found recurrently in our cohort were selected for validation. Recurrent mutations in *DNMT3A* leading to DNMT3A R882C and R882H were found to eliminate a recognition site for the restriction enzyme TauI [61]. Mutations in this residue were identified in 13 samples at different variant allele frequencies. These mutations were validated by PCR amplification (forward primer 5′-GAACTAAGCAGGCGTCAGAGGA-3′, reverse primer 5′-AAAAAGGGAAGGGGGAGGAAGG-3′) of a region of 577 bp surrounding the region of interest in *DNMT3A*, followed by restriction digestion. Amplicons of the wild-type sequence are digested by TauI in two fragments of 369 and 208 bp, while amplicons of either mutant allele fail to be recognized by the enzyme and remain undigested. Size analysis of the digested products was performed on a TapeStation. A subset of the digested products was selected for subsequent sequencing on an Ion Torrent platform.

References

1. Yadav, V. K., DeGregori, J. & De, S. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Res* **44,** 2075–2084 (2016).
2. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N Engl J Med* **371,** 2477–2487 (2014).
3. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N Engl J Med* **371,** 2488–2498 (2014).
4. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20,** 1472–1478 (2014).
5. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep* **10,** 1239–1245 (2015).
6. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat Commun* **7,** 1–7 (2016).
7. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).
8. Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492,** 438–442 (2012).
9. Gao, Z., Wyman, Mi. J., Sella, G. & Przeworski, M. Interpreting the Dependence of Mutation Rates on Age and Time. *PLOS Biol* **14,** e1002355 (2016).
10. Ramsey, M. J. *et al.* The effects of age and lifestyle factors on the accumulation of cytogenetic damage as measured by chromosome painting. *Mutat Res DNAging* **338,** 95–106 (1995).
11. McKerrell, T. & Vassiliou, G. S. Aging as a driver of leukemogenesis. *Sci Transl Med* **7,** 306fs38 (2015).
12. Stratton, M., Campbell, P. & Futreal, A. The cancer genome. *Nature* **458,** 719–724 (2009).
13. Adams, P. D., Jasper, H. & Rudolph, K. L. Aging-Induced Stem Cell Mutations as Drivers for Disease and Cancer. *Cell Stem Cell* **16,** 601–612 (2015).
14. Chaudhury, S. S., Morison, J. K., Gibson, B. E. S. & Keeshan, K. Insights into cell ontogeny, age, and acute myeloid leukemia. *Exp Hematol* **43,** 745–755 (2015).
15. Corces-Zimmerman, M. R. & Majeti, R. Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia* **28,** 2276–2282 (2014).
16. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126,** 9–16 (2015).
17. Akunuru, S. & Geiger, H. Aging, Clonality, and Rejuvenation of Hematopoietic Stem Cells. *Trends Mol Med* **22,** 701–712 (2016).
18. Beerman, I. *et al.* Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A* **107,** 5465–5470 (2010).
19. Geiger, H., de Haan, G. & Florian, M. C. The ageing haematopoietic stem cell compartment. *Nat Rev Immunol* **13,** 376–389 (2013).
20. Rozhok, A. I. & DeGregori, J. Toward an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations. *Proc Natl Acad Sci* **112,** 201501713 (2015).
21. Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res* **24,** 733–742 (2014).
22. Shlush, L. I., Zandi, S., Itzkovitz, S. & Schuh, A. C. Aging, clonal hematopoiesis and preleukemia: not just bad luck? *Int J Hematol* **102,** 513–522 (2015).
23. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* **23,** 843–854 (2013).
24. Neveling, K. *et al.* BRCA Testing by Single-Molecule Molecular Inversion Probes. *Clin Chem* (2016).
25. Eijkelenboom, A. *et al.* Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-

Embedded Tissue Using Single Molecule Tags. *J Mol Diagnostics* **18,** 851–863 (2016).
26. Weren, R. D. A. *et al.* Novel BRCA1 and BRCA2 Tumor Test as Basis for Treatment Decisions and Referral for Genetic Counselling of Patients with Ovarian Carcinomas. *Hum Mutat* n/a-n/a (2016). doi:10.1002/humu.23137
27. Carlston, C. M. *et al.* Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. *Hum Mutat* **84112,** (2017).
28. Goriely, A. & Wilkie, A. O. M. Paternal Age Effect Mutations and Selfish Spermatogonial Selection: Causes and Consequences for Human Disease. *Am J Hum Genet* **90,** 175–200 (2012).
29. Maher, G. J. *et al.* Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proc Natl Acad Sci* **113,** 2454–2459 (2016).
30. Yoon, S.-R. *et al.* Age-Dependent Germline Mosaicism of the Most Common Noonan Syndrome Mutation Shows the Signature of Germline Selection. *Am J Hum Genet* **92,** 917–926 (2013).
31. Giannoulatou, E. *et al.* Contributions of intrinsic mutation rate and selfish selection to levels of de novo HRAS mutations in the paternal germline. *Proc Natl Acad Sci* **110,** 20152–20157 (2013).
32. Goriely, A. *et al.* Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. *Nat Genet* **41,** 1247–1252 (2009).
33. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385,** 1305–1314 (2015).
34. Goriely, A., McVean, G. A. T., Röjmyr, M., Ingemarsson, B. & Wilkie, A. O. M. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* **301,** 643–6 (2003).
35. Choi, S.-K., Yoon, S.-R., Calabrese, P. & Arnheim, N. Positive Selection for New Disease Mutations in the Human Germline: Evidence from the Heritable Cancer Syndrome Multiple Endocrine Neoplasia Type 2B. *PLoS Genet* **8,** e1002420 (2012).
36. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat Neurosci* **17,** 764–772 (2014).
37. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* **19,** 1194–1196 (2016).
38. McRae, J. F. *et al.* Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv* 1–39 (2016).
39. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
40. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).
41. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (80-)* **355,** 752–756 (2017).
42. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47,** 1402–1407 (2015).
43. Moehrle, B. M. & Geiger, H. Aging of hematopoietic stem cells: DNA damage and mutations? *Exp Hematol* **44,** 895–901 (2016).
44. Mason, C. C. *et al.* Age-related mutations and chronic myelomonocytic leukemia. *Leukemia* (2015). doi:10.1038/leu.2015.337
45. Link, D. C. & Walter, M. J. 'CHIP'ping away at clonal hematopoiesis. *Leukemia* **30,** 1633–1635 (2016).
46. Grimwade, D., Ivey, A. & Huntly, B. J. P. Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance. **127,** 29–42 (2015).
47. Forbes, S. A. *et al.* COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43,** D805–D811 (2015).
48. Yang, L., Rau, R. & Goodell, M. A. DNMT3A in haematological malignancies. *Nat Rev Cancer* **15,** (2015).
49. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* **43,** 729–731 (2011).
50. Kosaki, R., Terashima, H., Kubota, M. & Kosaki, K. Acute myeloid leukemia-associated *DNMT3A* p.Arg882His mutation in a patient with Tatton-Brown-Rahman overgrowth syndrome as a

constitutional mutation. *Am J Med Genet Part A* **173,** 250–253 (2017).
51. Martinelli, S. *et al.* Heterozygous germline mutations in the CBL tumor-suppressor gene cause a noonan syndrome-like phenotype. *Am J Hum Genet* **87,** 250–257 (2010).
52. Niemeyer, C. M. *et al.* Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat Genet* **42,** 794–800 (2010).
53. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol* **34,** 531–8 (2016).
54. Galesloot, T. E. *et al.* Cohort Profile: The Nijmegen Biomedical Study (NBS). *Int J Epidemiol* dyw268 (2017). doi:10.1093/ije/dyw268
55. Champion, K. J. *et al.* Germline mutation in BRAF codon 600 is compatible with human development: De novo p.V600G mutation identified in a patient with CFC syndrome. *Clin Genet* **79,** 468–474 (2011).
56. Niihori, T. *et al.* Germline KRAS and BRAF mutations in cardio-facio-cutaneous syndrome. *Nat Genet* **38,** 294–6 (2006).
57. Tartaglia, M. *et al.* Diversity and functional consequences of germline and somatic PTPN11 mutations in human disease. *Am J Hum Genet* **78,** 279–90 (2006).
58. Groesser, L. *et al.* Postzygotic HRAS and KRAS mutations cause nevus sebaceous and Schimmelpenning syndrome. *Nat Genet* **44,** 783–787 (2012).
59. Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: Optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30,** 2670–2672 (2014).
60. O'Roak, B. J. *et al.* Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science* **338,** 1619–1622 (2012).
61. Mancini, M. *et al.* Two Novel Methods for Rapid Detection and Quantification of DNMT3A R882 Mutations in Acute Myeloid Leukemia. *J Mol Diagn* **17,** 179–184 (2014).

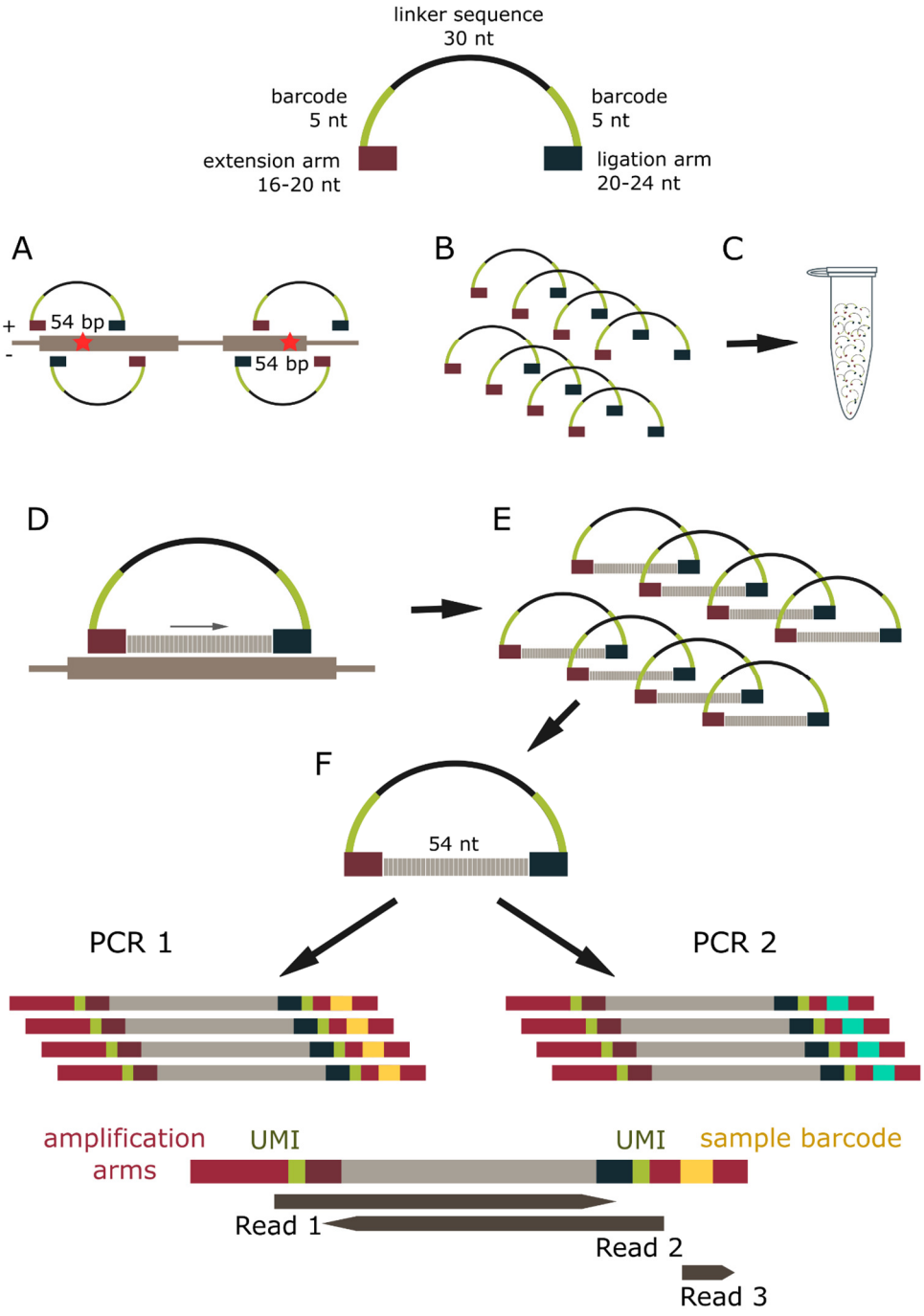
Supplementary data

Supplementary methods

Preparation of smMIP library

Individual smMIPs were pooled equimolarly and phosphorylated using T4 polynucleotide kinase and 10x T4 DNA ligase buffer supplemented with 10mM ATP (New England Biolabs). The smMIP capture was performed on 8µl of input DNA (200 ng) supplied with 17uL of capture mixture (0.01µl Ampligase DNA ligase (100U/µl, Illumina), 2.5µl 10x ampligase buffer (Illumina), 0.28µl phosphorylated smMIP pool dilution (corresponding to a DNA:smMIP ratio of 1:8000), 0.32µl Hemo Klentaq (10U/µl, New England Biolabs), 0.03µl dNTPs (0.25mM) and 13.86µl H2O). The capture mix was incubated for 18-22 hours at 60°C, after which the mix was cooled and treated with exonuclease (0.5µl Exonuclease I (New England Biolabs), 0.5µl Exonuclease III (New England Biolabs), 0.2µl 10x Ampligase buffer (Illumina) and 0.8µl H2O for 45 minutes at 37°C and 2 minutes at 95°C to inactivate the exonucleases). Each exo-treated sample was split in two technical replicates of 10µl, which were then amplified and barcoded by PCR independently (1.25µl of barcoded reverse primer (10µM), 12.5µl 2x iProof (BioRad Laboratories), 0.125µl forward primer (100µM) and 1.8µl H2O). The PCR products were run on gel, pooled and purified using AMPureXP Beads (Agencourt).

▶ **Figure S1. Overview of smMIP protocol used in this study.** Each smMIP is a single stranded 80 nucleotide-long DNA molecule consisting of the extension and the ligation arm (shown in burgundy and blue, respectively) which together are 40 nucleotides long and are designed to be complementary to the targeted region. The two arms are connected by a 30 nucleotide-long linker sequence (in black). Each smMIP contains a unique molecule identifier (UMI) composed by 2 barcodes of random 5 nucleotide sequences (shown in green). The smMIPs are designed for double tiling of regions of interest containing mutations (shown as red star) on the plus and the minus strand (A). The smMIPs are ordered as long oligonucleotides (B), after which they are pooled and phosphorylated (C). DNA capture is performed by mixing the phosphorylated smMIP probes with the DNA, dNTPs, polymerase and ligase to form the reverse complement of the region of interest to which the probe binds and ligate the probe into a circular single strand of DNA (D). Afterwards, the mix is digested with exonuclease to remove all linear DNA molecules (E). We subsequently separated the captured and circularized molecules in two separate technical replicates and performed PCR amplification separately (F), using a sample and PCR-specific barcode (shown in yellow and cyan). The PCR products were purified using AmpureXP beads and sequences on an Illumina Nextseq platform using 2×79 bp reads.


linker sequence
30 nt
barcode
5 nt
barcode
5 nt
extension arm
16-20 nt
ligation arm
20-24 nt
A
B
C
+
-
54 bp
54 bp
D
E
F
54 nt
PCR 1
PCR 2
amplification
arms
UMI
UMI
sample barcode
Read 1
Read 2
Read 3

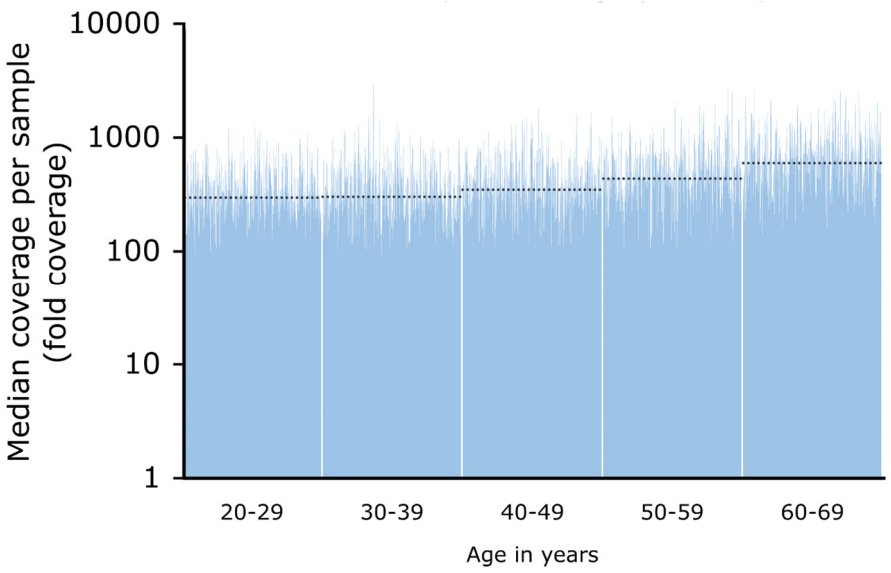
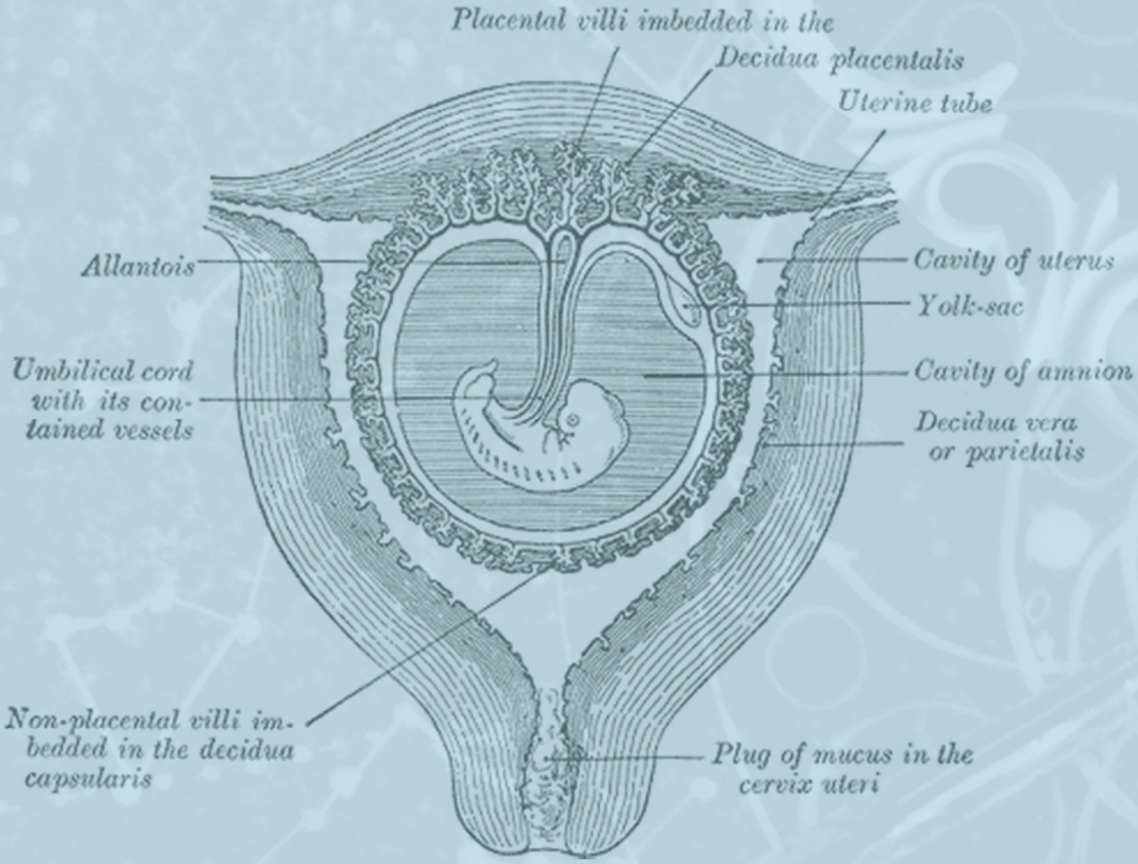


Figure S2. Median unique sequencing coverage per sample, corresponding to the number of unique DNA molecules sequenced per sample and per position after removal of PCR duplicates. Each sample is represented here by two bars, corresponding to each technical replicate. The median unique sequencing coverage per age group is shown as a horizontal dotted line. The overall median unique sequencing coverage was 418-fold per replicate and 845-fold per sample.

	Mean age	Median age	Total	Included	Male	Male (%)	Female	Female (%)
20-29	24.7	25	400	396	195	49.5	201	50.5
30-39	34.6	35	405	402	202	50.2	200	49.8
40-49	44.7	45	404	404	202	50.0	202	50.0
50-59	54.3	55	403	402	202	50.2	200	49.8
60-69	64.5	64	402	402	201	50	201	50.0
Total	44.6	45	2014	2006	1002	50.0	1004	50.0

Table S1. Age and sex of participants in our study.



Sectional plan of the gravid uterus in the third and fourth month.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 6:
Overlapping *SETBP1* gain-of-function mutations in Schinzel-Giedion syndrome and hematologic malignancies

Published as:

Acuna-Hidalgo R.*, Deriziotis P.*, Steehouwer M., Gilissen C., Graham S.A., van Dam S., Hoover-Fong J., Telegrafi A.B., Destree A., Smigiel R., Lambie L.A., Kayserili H., Altunoglu U., Lapi E., Uzielli M.L., Aracena M., Nur B.G., Mihci E., Moreira L.M., Borges Ferreira V., Horovitz D.D., da Rocha K.M., Jezela-Stanek A., Brooks A., Reutter H. Cohen J.S., Fatemi A., Smitka M., Grebe T., Di Donato N., Deshpande C., Vandersteen A., Marques Lourenço C., Dufke A., Rossier E., Andre G., Baumer A., Spencer C., McGaughran J., Franke L., Veltman J.A., De Vries B.B.A., Schinzel A., Fisher S.E., Hoischen A. and van Bon B.W. Overlapping *SETBP1* gain-of-function mutations in Schinzel-Giedion syndrome and hematologic malignancies. PLoS Genetics 2017 Mar 27;13(3):e1006683.

* These authors contributed equally to this work

Abstract

Schinzel-Giedion syndrome (SGS) is a rare developmental disorder characterized by multiple malformations, severe neurological alterations and increased risk of malignancy. SGS is caused by *de novo* germline mutations clustering to a 12 bp hotspot in exon 4 of *SETBP1.* Mutations in this hotspot disrupt a degron, a signal for the regulation of protein degradation, and lead to the accumulation of SETBP1 protein. Overlapping *SETBP1* hotspot mutations have been observed recurrently as somatic events in leukemia.

We collected clinical information of 47 SGS patients (including 26 novel cases) with germline *SETBP1* mutations and of four individuals with a milder phenotype caused by *de novo* germline mutations adjacent to the *SETBP1* hotspot. Different mutations within and around the *SETBP1* hotspot have varying effects on SETBP1 stability and protein levels *in vitro* and in *in silico* modeling. Substitutions in SETBP1 residue I871 result in a weak increase in protein levels and mutations affecting this residue are significantly more frequent in SGS than in leukemia. On the other hand, substitutions in residue D868 lead to the largest increase in protein levels. Individuals with germline mutations affecting D868 have enhanced cell proliferation *in vitro* and higher incidence of cancer compared to patients with other germline *SETBP1* mutations.

Our findings substantiate that, despite their overlap, somatic *SETBP1* mutations driving malignancy are more disruptive to the degron than germline *SETBP1* mutations causing SGS. Additionally, this suggests that the functional threshold for the development of cancer driven by the disruption of the SETBP1 degron is higher than for the alteration in prenatal development in SGS. Drawing on previous studies of somatic *SETBP1* mutations in leukemia, our results reveal a genotype-phenotype correlation in germline *SETBP1* mutations spanning a molecular, cellular and clinical phenotype.

Introduction

Schinzel-Giedion syndrome (SGS; OMIM 269150) is a rare developmental disorder characterized by multiple malformations including midface hypoplasia, cardiac defects, hydronephrosis and skeletal abnormalities.[1–3] This clinically recognizable syndrome was the first dominant disorder for which the underlying genetic cause was identified by whole exome sequencing.[4] In 12 of 13 unrelated individuals with this disorder, we identified germline *de novo* mutations in *SETBP1* clustering to a hotspot of 12 base pairs coding for residues 868 to 871 of the SETBP1 protein.[4] Interestingly, shortly after the identification of germline *de novo* mutations in *SETBP1* as the cause of SGS, overlapping somatic mutations in *SETBP1* were reported in several types of myeloid malignancies.[5–7] This dual role in cancer and development is not unique to *SETBP1;* a growing number of genes in which germline mutations cause developmental disorders, such as *HRAS*, *ASXL1*, *EZH2* and *FGFR2*, are also known to harbor overlapping somatic mutations which drive the development of cancer.[8] This genetic overlap is not entirely unexpected; higher rates of childhood cancer have been identified in individuals with birth defects and *vice versa*,[9–11] a finding which is thought to be the consequence of abnormalities in molecular pathways shared between embryogenesis and cancer development.[12,13]

The precise function of *SETBP1*, which encodes the SET-binding protein 1 (OMIM 611060), is yet to be discovered and, as a result, the molecular consequences of *SETBP1* mutations remain largely unknown. However, the clustering of all germline *SETBP1* mutations identified in SGS to a single region and their overlap with the somatic events identified in myeloid malignancies support a gain-of-function effect on the SETBP1 protein. This recurrently mutated region of the protein is highly conserved and has been identified as a degron signal targeted by the SCF-βTrCP1 E3 ligase.[5] A degron is a peptide sequence that is recognized and bound by a component of the ubiquitin-proteasome pathway, thereby initiating degradation of the protein by ubiquitination.[14] As a result, mutations localizing to the degron in SETBP1 disrupt binding by the βTrCP1 E3 ligase, increase protein stability by interfering with ubiquitination[15] and ultimately lead to accumulation of SETBP1 protein in cells.[5] While the molecular consequences of germline *SETBP1* mutations are poorly understood, somatic

mutations disrupting the SETBP1 degron lead to increased proliferation in myeloid progenitors,[7] possibly mediated by effects on its interaction partner SET, phosphorylation of PP2A and transcriptional activation of *HOXA9* and *HOXA10*.[5,16,17]

Additional clinical and functional investigation is warranted to gain more understanding about the molecular mechanisms of SGS. Here we present the clinical characterization of the largest cohort of individuals with genetically confirmed SGS and establish genotype-phenotype correlations for SGS. Given the occurrence of overlapping germline and somatic *SETBP1* mutations, we compare the mutations in SGS and leukemia to identify genetic and functional differences between *SETBP1* mutations in both conditions.

Results

Clinical features of SGS

Classic SGS is caused by mutations within four residues of the SETBP1 degron (D868, S869, G870 and I871, Figure 1A and 1B), constituting the critical consensus sequence of the degradation signal[18] (from here on, referred to as the canonical degron). Since the initial description of 12 mutation-positive cases,[4] we have gathered clinical details of 26 additional individuals with SGS genetically confirmed by the presence of *de novo* mutations in *SETBP1* (see Table 1). In addition, we present three patients with a milder phenotype variably overlapping with SGS and secondary to novel mutations in *SETBP1* affecting highly conserved residues in close proximity to the canonical degron (current cases no. 27-29, with mutations in residues E862, S867 and T873 shown in green in Figure 1B). We report on the clinical features observed in our cohort in addition to previously published mutation-positive cases to further delineate this disorder (n=51 individuals).

Dysmorphic facial features

All individuals with SGS have characteristic facial features that are easily recognizable (Figure 1C and Supplementary Figure S1). Typically, SGS patients have large fontanelles (n=37/41), a prominent forehead (n=41/44), bitemporal narrowing (n=32/36), shallow orbits or prominent eyes (n=36/38), hypertelorism (n=36/42) a retracted and shortened midface (n=45/45) and full cheeks, leading to a facial frontal silhouette in the shape of a number eight.[19] Additionally, most patients have a deep groove under the eyes (n=38/39), upslanting palpebral fissures (n=21/26) and a short nose with a bulbous nasal tip (n=43/44). Some individuals may present with less recognizable facial features in the first weeks after birth and after the age of 18 months. In those cases, additional diagnostic clues that may facilitate the diagnosis include the abnormal shape of the ears

(n=38/39). Classically, the ears of individuals with SGS are low-set and posteriorly rotated with anteriorly angulated lobules giving rise to a question mark shape (Figure 1D). In individuals who do not have the typical lobules, the majority does have folded helices and prominent anti-helices. About half of the patients show a large mouth (n=17/33) with an everted lower lip and a protruding tongue (n=18/38). In addition, they may present with micrognathia (n=29/30) and a philtrum groove. Hypertrichosis was identified in two thirds of the patients (n=26/37), facial hemangioma in eight of 33 cases and a short neck in 30 of 33 cases.

Skeletal features

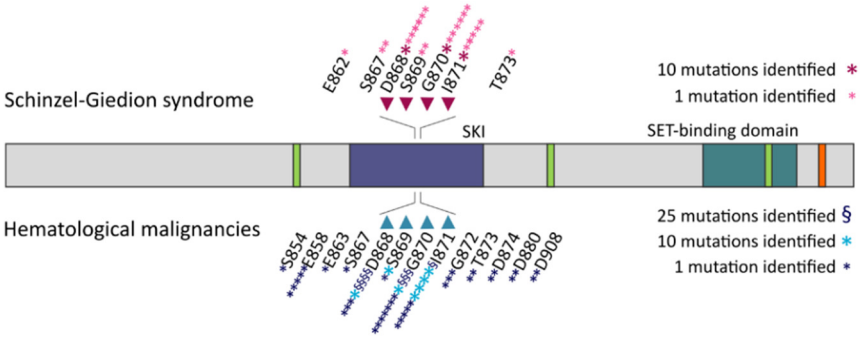
As molecular testing for *SETBP1* mutations has become available, the extent of diagnostic radiologic evaluations has diminished and, therefore, has not been performed in all patients within this cohort (data were available for 31 cases). Skeletal characteristics present in over 75% of these patients included a sclerotic base of the skull (n=19/23) with wide occipital synchondrosis (n=18/23), widening of ribs (n=27/31), short pubic rami, wide pubic symphysis (n=18/20) and hypoplastic distal phalanges in hands and feet (n=21/25). In case 22, the synchondrosis had closed completely in a later radiologic examination. Post-axial polydactyly was noted in only 11% of patients (n=4/38). Retrospective study of photographs of the hands of individuals with SGS shows that a typical posture with clenched fingers is common (Figure 1E and 1F).

Neurologic features

Microcephaly was observed in approximately two thirds of the individuals with SGS (n=29/39). The occipitofrontal circumference in the remaining individuals for whom data were available, was always below the 50th percentile and often near the 10th percentile. Severe developmental delay occurs in all individuals with SGS and 95% present with epilepsy (n=42/44). Almost all common types of epilepsy occur and seizures are characteristically extremely refractory to treatment with medication or ketogenic diet. Many patients have hearing (n=24/27) or vision impairment (n=20/26) thought to be of cerebral origin. Spasticity was noted in 17 out of 20 of cases.

Structural brain abnormalities are variable in SGS. The most common anomaly is hypoplasia or aplasia of the corpus callosum (n=31/38). Additional abnormalities often encountered are cortical atrophy (n=18/33), ventricle anomalies (n=26/42), abnormal gyration and delayed myelination and choroid plexus cysts (n=13/31).

A



B

DpSGφXpS/T

H. sapiens	PVSESHSEETIPS**DSGIGT**DNNSTSDQAEKSS
P. troglodytes	PVSESHSEETIPS**DSGIGT**DNNSTSDQAEKSS
M. musculus	PVSESHSEETIPS**DSGIGT**DNNSTSDQAEKSS
G. gallus	PVSESHSEETIPS**DSGIGT**DNNSTSDQAEKSS
X. tropicalis	PISESHSEETIPS**DSGIGT**DNNSTSDQAEKSS
D. rerio	PVSESHSEETIPS**DSGIGT**DNNSTSDQTEKGP

◀ **Figure 1. Genetic and clinical characteristics of individuals with germline *SETBP1* mutations and Schinzel-Giedion syndrome. A.** Schematic representation of the SETBP1 protein, indicating changes found in SGS and in hematologicAL malignancies. The residues of the canonical degron are highlighted with arrows. Protein domains of SETBP1 are shown in different colors with green corresponding to three AT hooks, purple to the SKI homologous region, blue to the SET binding domain and orange to a repeat domain (modified from Piazza *et al.*[5]). **B.** Sequence alignment of the region containing the degron of SETBP1 (in bold) in human (Uniprot accession number Q9Y6X0), chimpanzee (H2QEG8), mouse (Q9Z180), chicken (E1BWZ2), african clawed frog (F6TBV9) and zebrafish (B0R147). The consensus motif for βTrCP1 substrates is shown on top, with ϕ representing a hydrophobic residue and X any amino acid. Residues in which pathogenic germline mutations have been identified in classic SGS are highlighted in blue, while residues in which novel mutations leading to an atypical form of SGS are shown in green. **C.** Distinctive facial features encountered in classic SGS (current case 9 at 1,5 years of age). **D.** Typical question mark-shaped ear from observed in current case 18. **E.** Characteristic hand posture with clenched fingers, observed from current case 16. **F.** Facial features of current case 27 with a mutation in SETBP1 residue S867 at 4 years of age. Note the clenched fingers. **G.** Facial features of current case 28 with a mutation in SETBP1 residue E862 at 5 years of age. **H.** Facial features of current case 29 with a mutation in SETBP1 residue T873 at the age of 23 months.

Additional congenital anomalies

Individuals with SGS nearly always present hydronephrosis (n=45/47), a feature that can be detected during routine prenatal medical exams. Two patients in our cohort were noted to have hydronephrosis at 20 weeks of gestation. Other kidney anomalies include abnormal ureters, renal cysts and stones. Almost all patients have genital anomalies (n=41/45), which include hypospadias, underdeveloped genitalia and displaced anus. Half of the patients (n=20/43) have structural cardiac malformations, the majority of which present with defects of the atrial septum. Other anomalies include patent foramen ovale, patent ductus arteriosus and cardiac hypertrophy. Alterations in the internal organs may be identified in some individuals with SGS, including hypoplasia of the pancreatic tail or hepatosplenomegaly. Microscopic evaluation in one patient (current patient no. 6) showed dilated glands and mucus depositions in intra-acinar pancreatic ducts at a post-mortem examination at 4 days of age. The observed features were similar to the mucus obstruction seen in cystic fibrosis. However, *CFTR* analysis in this patient proved negative. Nineteen patients (n=19/27) were noted to have alacrima. Inguinal hernia (n=8/15) and talipe(s) equinovarus (n=11/17) were also frequently noted.

Swallowing and breathing difficulties

A major medical problem encountered in the care of individuals with SGS is the difficulty in swallowing and breathing. This is caused by a combination of factors such as structural abnormalities of the respiratory apparatus (*e.g.* choanal

Residue affected in SETBP1	E862	S867	D868	S869	G870	I871	T873	All degron affecting mutations (868-871)	
Male(M):female(F)	1F	2F	8F:7M	2F	5F:10M	6F:9M	1M	21F:26M	
Microcephaly	1/1	1/2	10/12	1/2	10/13	8/11	0/1	29/39	74.4%
SGS facial gestalt	0/1	2/2	15/15	2/2	15/15	15/15	0/1	47/47	100%
Hydronephrosis	0/1	0/2	15/15	2/2	14/15	14/15	0/1	45/47	95.7%
Genital abnormalities	1/1	0/2	14/15	1/2	14/15	12/13	0/1	41/45	91.1%
Cardiac defects	0/1	1/2	10/15	1/2	4/13	5/13	0/1	20/43	46.5%
Tracheo- or laryngomalacia	0/1	0/1	3/4	0/2	3/8	2/2	0/1	8/16	50.0%
Inguinal hernia	0/1	0/1	2/4	0/2	6/8	0/1	0/1	8/15	53.3%
Alacrima	0/1	2/2	6/10	0/2	7/9	6/6	0/1	19/27	70.4%
Developmental delay	1/1	2/2	14/14	2/2	13/13	10/10	1/1	39/39	100%
Seizures	0/1	2/2	15/15	2/2	13/14	12/13	0/1	42/44	95.5%
Spasticity and/or hypertonia	0/1	1/1	4/4	1/2	8/10	4/4	1/1	17/20	85.0%
Vision impairment	1/1	1/1	7/10	1/2	7/8	5/6	0/1	20/26	76.9%
Hearing impairment	0/1	0/1	9/9	0/1	7/8	8/9	0/1	24/27	88.9%
Progressive failure to thrive	0/1	0/1	10/11	1/2	13/13	8/9	0/1	32/35	91.4%
Ventriculomegaly	0/1	NA	6/12	2/2	11/14	7/14	0/1	26/42	61.9%
Underdeveloped corpus callosum	0/1	NA	9/11	0/2	12/13	10/12	0/1	31/38	81.6%
Cortical atrophy or dysplasia	0/1	NA	8/10	0/2	7/10	3/11	1/1	18/33	54.5%
Choroid plexus cysts	0/1	NA	2/9	0/2	8/10	3/10	0/1	13/31	41.9%
Sclerotic base of skull or mastoid	NA	NA	9/10	0/1	5/5	5/7	NA	19/23	82.6%
Hypoplastic distal phalanges	0/1	NA	8/9	0/1	8/9	5/6	NA	21/25	84.0%
Broad ribs	0/1	NA	10/13	2/2	6/7	9/9	NA	27/31	87.1%
Hypoplastic pubic bones	0/1	NA	6/7	2/2	4/5	6/6	NA	18/20	90.0%
Tumors	0/1	0/2	5/11	0/2	1/11	1/9	0/1	7/33	21.2%

Table 1. Major clinical findings in 51 individuals with germline mutations in *SETBP1*. NA stands for "Not Assessed".

stenosis, n=10/34), tracheobronchomalacia (n=8/16), lung hypoplasia, poor management of oral and respiratory secretions (*e.g.* resulting from micrognathia, gingiva hypertrophy or excessive mucus production) and a high susceptibility to airway infections. Although gingiva hypertrophy can result from the use of anti-epileptic medication, this feature was already noted in a newborn with SGS at four days of age (patient no.6). In this case, dilated laryngeal and broncheal glands filled with mucus were observed in the microscope, consistent with a previous report of thickened alveolar mucosa and fibrous hyperplasia of the gingiva with mucoid depositions,[20] suggesting that this is a feature of SGS.

Cause of death

Most affected individuals do not survive past childhood, with pneumonia as a major cause of death in SGS (n=8/15). Other reported causes in early infancy include congenital cardiac defects, tumors, lung hypoplasia, intractable seizures and sudden cardiac arrest. Six individuals developed solid tumors, predominantly of neuroepithelial origin in the lumbosacral region. Additionally, one individual in this cohort developed juvenile myelomonocytic leukemia. Although the majority of individuals die during infancy, five out of twelve patients with a protein substitution in residue G870 lived to the age of 5 or older (5 to 15 years of age). The average age of death of deceased individuals with a substitution in D868, S869, G870 and I871 is 18 months (n=10), 32 months (n=2), 48 months (n=7) and 25 months (n=8), respectively (see Supplementary Figure S3). Due to the small size of the cohort, it is not possible to draw conclusions on whether there are differences in survival depending on the mutated residue.

Individuals with mutations outside of the degron

Individuals with germline *SETBP1* mutations occurring outside the degron (n=4) are reported in this separate section. Both patients with a mutation in residue S867 (one reported in reference 21 and current case no. 25 from our cohort, see Figure 1G and Figure S2A-D) had a characteristic facial appearance, genital anomalies and seizures but did not show hydronephrosis. Other features of these patients fit within the spectrum of SGS and are summarized in Table 1.

One individual with a mutation in residue E862 did not have a characteristic facial appearance (Figure 1H and Figure S2E-H), seizures nor hydronephrosis. However, this patient had several other overlapping features with SGS, including dysphagia requiring a gastrostomy tube, vision loss due to retinal dystrophy, bilateral renal cysts and severe spasticity. She showed microcephaly, prognathism, small feet with short toes and a normal height. At five years of age, she had severe intellectual disability and could neither speak nor walk.

The individual with a mutation in residue T873 had the mildest phenotype, presenting with developmental delay, autistic features, spastic diplegia and milder dysmorphic features (Figure 1I and Figure S2I-L). This patient does not present any of the major congenital anomalies commonly found in SGS and, despite developmental delay, his initial developmental outcome seemed higher compared to individuals with *SETBP1* mutations within the degron. He achieved several milestones: smiling at eight weeks, making vowel sounds at 14 months, sitting unassisted at 22 months and, at 2 years of age, he was able to maintain crawling position, walk with help and feed himself. Thereafter, he entered a regression phase with loss of the aforementioned milestones and

started self-injurious behavior. An IQ test was not available but, at 4 years of age, his functional level was estimated to be that of an eight-month-old.

Functional characterization of SETBP1 variants

Mutations within the canonical degron of SETBP1 disrupt its interaction with βTrCP1, increasing protein stability and SETBP1 levels[5,15] but the effect of *SETBP1* mutations outside the canonical degron on protein stability is unknown. Clear differences in the severity of the phenotype resulting from germline *SETBP1* mutations within and outside the canonical degron suggest a difference in the effect of the mutations depending on their localization. Therefore, we performed the functional characterization of the most frequent *SETBP1* mutations observed within the canonical degron and causative for classic SGS (D868N, S869N, G870S and I871T) as well as one mutation outside the canonical degron leading to atypical SGS (E862K).

Effects of SETBP1 variants on protein levels and protein stability

To investigate the effect of SETBP1 variants on protein stability, we quantified expression levels of YFP-fusion proteins in live HEK293 cells based on fluorescence intensity (Supplementary Figure S4). All three pathogenic SETBP1 variants E862K, D868N and I871T showed increased protein levels compared to WT SETBP1 (Figure 2A; $p<0.001$, ANOVA). Interestingly, SETBP1 variants occurring within the canonical degron (D868N, I871T) showed higher protein levels compared to the variant adjacent to the degron sequence (E862K; $p<0.001$ versus other mutants, ANOVA).

To verify that the increased protein levels seen in cells expressing pathogenic SETBP1 variants are due to resistance to degradation by the proteasome, we treated transfected HEK293 cells with MG132, a proteasome inhibitor (Figure 2B). WT SETBP1 protein levels were sensitive to inhibition of the proteasome, as evidenced by a significant increase in fluorescence intensity after MG132 treatment for 3 hours (132% vs 108% for treated versus untreated cells compared to baseline, $p<0.05$, Student's T-test). The difference in expression levels after 3 hours of MG132 treatment was not significant for SETBP1 variants E862K, D868N and I871T (E862K: 126% vs 113%; D868N: 118% vs 115%; I871T: 121% vs 121% for treated versus untreated cells compared to baseline). Prolonged treatment with MG132 resulted in a larger fold increase in fluorescence observed for treated versus untreated cells in cells transfected with WT versus mutant SETBP1 (1.58 for WT, 1.32 for E862K, 1.14 for D868N and 1.21 for I871T; $p<0.01$, ANOVA). Indeed, prolonged MG132 treatment results in a significant increase in expression levels of wild-type SETBP1 and the E862K mutant (SETBP1:

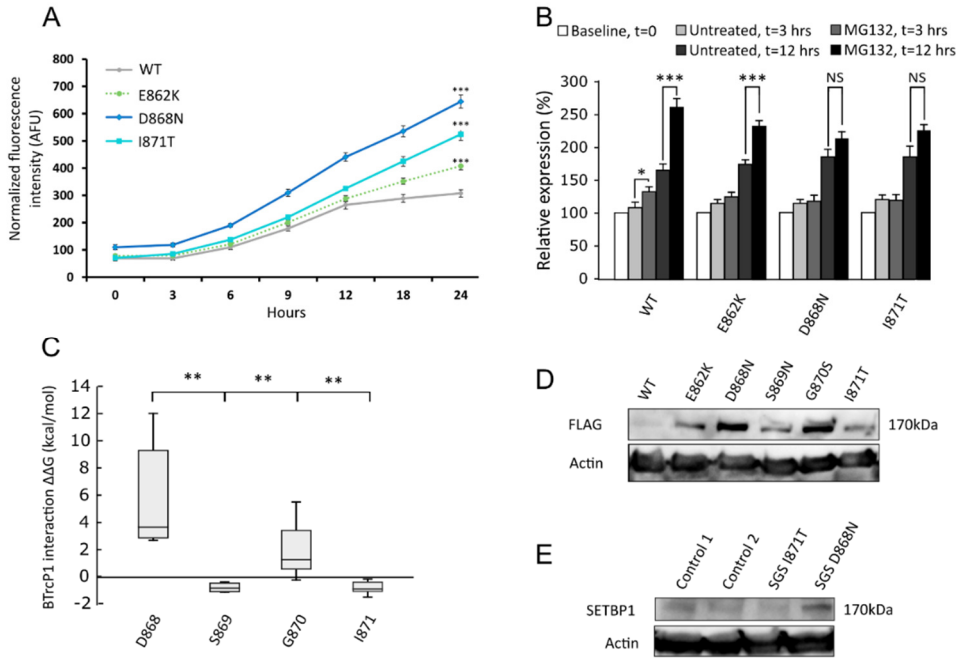


Figure 2. Functional analysis of SETBP1 mutations identified in SGS. A. Fluorescence measurements in live HEK293 cells expressing YFP-tagged SETBP1 variants. (*** $p<0.001$ versus wild-type and all mutants, ANOVA). All SETBP1 mutations studied displayed a statistically significant difference compared to wild-type and to all other mutations. This graph is representative of 3 independent experiments performed, with 6 technical replicates per experiment. Bars represent the standard error. **B.** Relative expression of SETBP1 protein variants in live HEK293 cells treated with MG132 proteasome inhibitor or vehicle only. Bars represent the standard error. (*** $p<0.001$, * $p<0.05$, NS: not significant, Student's T test and Mann-Whitney U test). **C.** ΔΔG values for degron-βTrCP1 interaction for all germline mutations reported in SETBP1 per residue (** $p<0.01$ D868 versus other residues; ANOVA). **D.** Immunoblot of whole cell lysates of HEK293 cells expressing FLAG-tagged SETBP1 variants probed with anti-FLAG antibody. **E.** Immunoblot of whole-cell lysates of fibroblasts probed with anti-SETBP1 antibody. Fibroblasts were derived from two cases of SGS, one carrying the I871T variant and the other carrying the D868N variant, as well as from two unrelated controls. In D and E, blots were stripped and re-probed with anti-β-actin antibody.

261% vs 164%; E862K: 230% vs 173% for treated versus untreated cells compared to baseline, $p<0.001$, Student's T test). Prolonged MG132 treatment did not significantly affect the expression levels of the D868N and I871T SETBP1 variants (211% vs 185% and 225% vs 186% for treated versus untreated cells compared to baseline, respectively). Cycloheximide chase of wild-type and mutant SETBP1 suggests decreased degradation of mutant SETBP1 as compared to the wild-type protein (Supplementary Figure S5), further supporting that mutations in the SETBP1 degron confer increased stability to the protein. This suggests that pathogenic SETBP1 variant proteins have decreased sensitivity or resistance to proteasome inhibition with varying magnitude of effects.

Different magnitude of effect of SETBP1 variants within the canonical degron

To explore the effect of disease-causing SETBP1 variants on the interaction of SETBP1 with βTrCP1, we performed *in silico* modeling of all known germline mutations that occur within the canonical degron (D868-I871) identified in patients with classic SGS. Although the protein structure of SETBP1 is not known, the sequence of the βTrCP1 binding site in the SETBP1 degron is similar to the sequence of a degron in β-catenin (Supplementary Table S1).[18] We therefore used a protein model of the degron of β-catenin in complex with βTrCP1 as a template to analyze the effect of pathogenic germline *SETBP1* variants on molecule stability and on the degron/βTrCP1 interaction energy (molecule ΔΔG and βTrCP1 interaction ΔΔG, respectively; Figure 2C and Supplementary Table S2). Modeling of all pathogenic germline substitutions observed in the degron shows that the interaction with βTrCP1 was most affected for substitutions of the aspartate residue from the motif ($p<0.001$ versus other residues, ANOVA), which is in line with previous findings.[18] This would entail that among the variants observed in SGS, mutations in residue D868 would have the largest effect on degron/βTrCP1 interaction, followed by mutations in G870.

To verify the differences on protein stability observed at the computational level, we used immunoblotting to examine the protein levels of SETBP1 variants in transfected HEK293 cells. Pathogenic SETBP1 variants led to substantially higher protein levels than WT SETBP1 (Figure 2D and Supplementary Figure S6). Variants D868N and G870S resulted in dramatically increased protein levels, whereas E862K, S869N and I871T showed a more modest increase in protein levels (Figure 2D). To further explore these findings, we performed immunoblotting to examine endogenous SETBP1 protein levels in fibroblasts derived from individuals with SGS. Fibroblasts from an individual carrying the D868N variant showed increased SETBP1 protein levels compared to two unrelated age-matched controls, whereas fibroblasts from a patient carrying the I871T mutation did not show any differences in endogenous SETBP1 levels compared to the controls (Figure 2E). Analysis of mRNA levels showed decreased *SETBP1* mRNA levels in cells of individuals with SGS as compared to controls, suggesting that the increase observed in endogenous SETBP1 protein does not result from increased *SETBP1* transcription (Supplementary Figure S7). Together, these results suggest that SETBP1 degron mutations have variable effects on protein stability, with the D868N variant having a stronger effect than I871T.

Overlapping SETBP1 mutations in SGS and in myeloid malignancies

Overlapping mutations in *SETBP1* have been identified as germline *de novo* events in SGS and as somatic mutations in myeloid malignancies (Figure 1A). Considering that canonical degron mutations vary in the magnitude of their effect, we compared germline and somatic *SETBP1* mutations to detect differences between the mutations in both conditions. In total, 48 germline *de*

novo mutations within the SETBP1 degron have been reported in individuals with SGS, both from our cohort and from published literature.[4,21–29] Similarly, we have retrieved from literature 245 individual somatic mutations in the SETBP1 degron, associated with different myeloproliferative disorders (see Supplementary table 4).[5–7,30–48] While the mutations overlap in both conditions, the distribution of the mutations within the canonical degron sequence is not the same; a significantly higher number of mutations affect residue I871 in SGS cases compared to myeloid malignancy cases (29% and 12%, respectively, $p<0.01$, Fisher's test with Bonferroni correction; see Figure 3A).

All SETBP1 degron mutations identified in leukemia were compared to mutations observed in SGS on their effect on the SETBP1-βTrCP1 interaction using *in silico* modeling (Figure 3B). Taking into account the frequency of each mutation, we then compared the difference in degron-βTrCP1 interaction energy for SETBP1 mutations observed in SGS versus those observed in myeloid malignancies (Figure 3C). Generally, mutations observed in myeloid malignancies showed higher ΔΔG values than mutations observed in SGS ($p <0.05$, Mann-Whitney's U test). However, the difference in ΔΔG between germline and somatic *SETBP1* mutations after exclusion of mutations in codon I871 is no longer statistically significant. This suggests that the difference between ΔΔG values between germline and somatic *SETBP1* mutations may be secondary mainly to the prevalence of mutations at codon I871 in each condition.

Similarities in downstream consequences of germline and somatic SETBP1 mutations

The SETBP1-SET interaction regulates SET protein levels and can induce cleavage of SET.[16] Higher levels of SET protein have been reported with overexpression of wild-type SETBP1 in HEK cells and of SETBP1 G870S in TF1 cells.[5,16] To establish whether pathogenic germline *SETBP1* degron mutations D868N, S869N and I871T also lead to increased SET protein levels in SGS, we performed immunoblotting on protein lysates of lymphoblastoid cell lines (LCLs) derived from three unrelated patients with SGS. Compared to age-matched controls, we observed that cell lines derived from individuals with germline mutations in the *SETBP1* degron show higher SET protein levels, with an increase in full length versus cleaved SET protein (Figure 4A and 4B).

Somatic mutations in *SETBP1* drive the development of myeloid malignancies by increasing proliferation in leukemic cells.[15] We examined LCLs derived from individuals with germline *SETBP1* mutations seen in SGS to determine whether they presented increased proliferation. In a time course experiment, we observed that LCLs derived from individuals with SGS proliferate

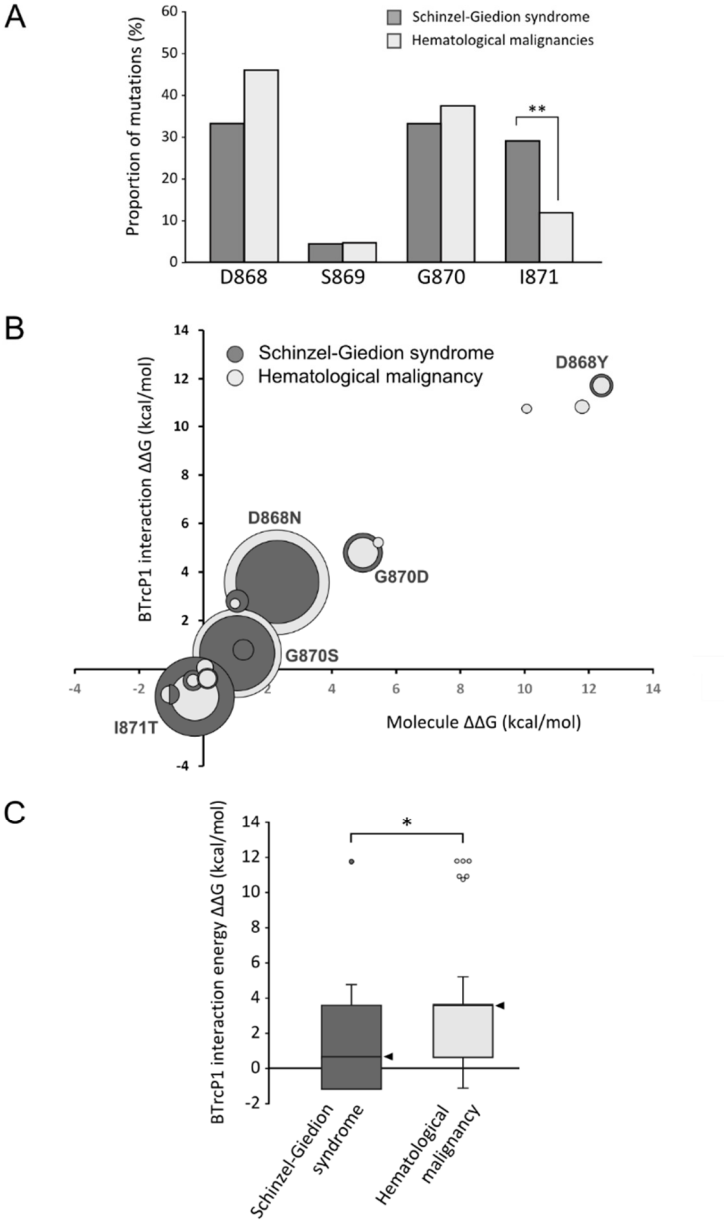


Figure 3. On average, *SETBP1* mutations seen in cancer are more severe than those observed in SGS. A. Distribution of mutations within the SETBP1 degron in SGS and in hematological malignancies. (** $p<0.01$, Fisher's test and Bonferroni correction for multiple testing). **B.** ΔΔG values for protein stability (x-axis) and degron-βTrCP1 interaction (y-axis) for all mutations reported in SETBP1. The size of each circle is proportional to the frequency of the mutation in each condition. **C.** Difference in free energy of binding in the interaction between βTrCP1 and the degron of variants arising from germline or somatic SETBP1 mutations compared to that of the interaction between βTrCP1 and the wild-type degron (* $p<0.05$, Mann-Whitney's U test). The median is highlighted by an arrow head.

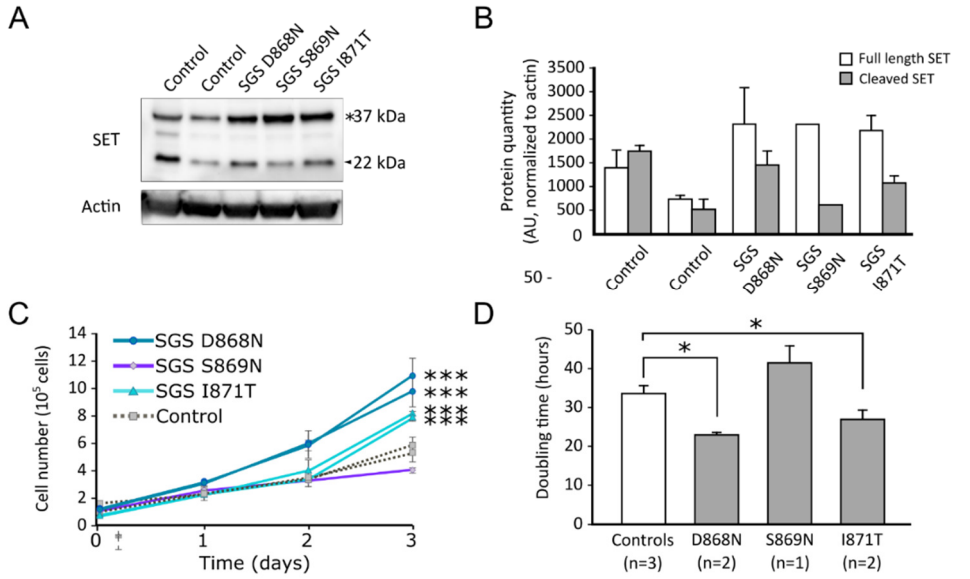


Figure 4. Increased SET protein and proliferation in cell lines of patients with SGS. A. Immunoblot of whole cell lysates of lymphoblastoid cell lines (LCLs) derived from three SGS patients and two age-matched controls. The blot was probed with anti-SET, before stripping and re-probing with β-actin to confirm equal loading. * denotes full-length SET protein (37kDa), arrowhead denotes cleaved SET protein (22kDa). **B.** SET protein levels as determined by densitometric analysis of Western blot data and normalized to actin (n=2 independent experiments). Bars represent the standard error. **C.** Proliferation assay of LCLs derived from SGS individuals and age-matched controls over 3 days (*** p< 0.001 versus controls, ANOVA). Bars represent the standard error. **D.** Mean doubling time for LCLs derived from individuals with germline *SETBP1* mutations compared to controls over 3 days (* p<0.05 versus controls, ANOVA). The mean of two experiments is shown, bars represent the standard error.

faster and have shorter doubling times than cells derived from age-matched controls in a genotype-dependent manner (p<0.001 versus controls by ANOVA, see Figure 4C and 4D). LCLs from unrelated individuals carrying D868N mutations had the shortest doubling times, followed by LCLs from unrelated patients carrying I871T mutations and by cells derived from age-matched controls (23.2, 27 and 33.8 hours respectively, p<0.05, ANOVA). LCLs from an individual with a S869N mutation consistently showed lower proliferation and longer doubling time (41.7 hours; Figure 4D). RNA sequencing of LCLs of individuals with germline *SETBP1* mutations revealed differential expression of 1,811 genes between controls and individuals with SGS (Supplementary Figure S8), of which 632 were upregulated and 1,179 were downregulated.[49] Gene set enrichment analysis of differentially expressed genes between controls and individuals with germline *SETBP1* mutations shows an enrichment for genes involved in mRNA transcription and translation, mitochondrial respiration and cell cycle (Supplementary Table S4).[50]

Increased tumorigenesis in individuals with mutations in residue D868

To determine whether a correlation exists between the magnitude of effect of a germline *SETBP1* mutation and tumorigenesis in individuals with SGS, we examined the clinical data of our cohort and previously published mutation-positive cases. In total, 7 malignancies have been reported in mutation-positive individuals, of which 5 occurred in patients with mutations in residue D868 (four tumors of primitive neuroectodermal origin and one myeloid leukemia). The remaining tumors included one ependymal tumor with myxopapillary and ependymoblastic differentiation in an individual with mutation G870D and one primitive neuroectodermal tumor in an individual with mutation I871T. To compare the intensity of effect of mutations in the group of individuals who developed cancer versus those who did not, we calculated the median SETBP1-βTrCP1 interaction ΔΔG value for each group based on results from *in silico* protein modeling (Figure 3B and Figure 5). The group of individuals who developed cancer carry *SETBP1* mutations which are more disruptive to the interaction of SETBP1 with βTrCP1 than the group of patients who did not develop cancer (Figure 5; βTrCP1 interaction ΔΔG = 3.57 vs 0.65, p = 0.029, Mann-Whitney U test). Finally, we analyzed the prevalence of malignancy per genotype in 33 mutation-positive individuals with SGS for whom we had data. While the incidence of tumorigenesis is 21.2% in this pooled group, individuals with mutation D868 have a statistically significantly higher risk of developing tumors than individuals with other germline mutations in *SETBP1* (OR = 9.16, 95% CI = 1.4 – 59.6, p = 0.02). Remarkably, mutations in residue D868 represent the most prevalent mutation observed somatically in cancer.

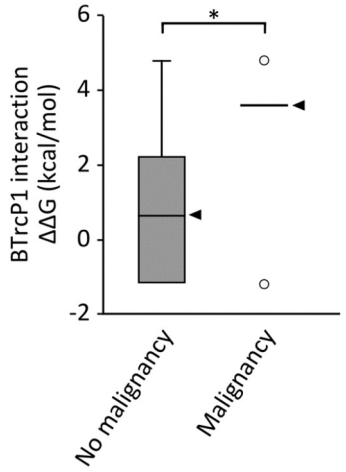


Figure 5. Correlation between functional effects of germline *SETBP1* mutations and risk of malignancy. Degron-βTrCP1 interaction ΔΔG for *SETBP1* mutations in individuals with SGS who did not develop a malignancy versus those who did (*p<0.05, Mann-Whitney's U test). The median for each group is marked by an arrowhead. The criteria to be considered negative for the development of a malignancy was either reaching the age of 60 months or dying without developing a malignant tumor or leukemia.

Discussion

The aims of our study were: to present the phenotype of the largest cohort of individuals with SGS and germline *SETBP1* mutations, establish genotype-phenotype correlations for germline *SETBP1* mutations and, by using these and previous findings from cancer research, provide insight into the molecular mechanism of SGS.

We present the clinical features of 47 individuals with "classic" SGS caused by germline mutations in *SETBP1* affecting the canonical degron (D868-I871). All mutations were found *de novo* in the affected proband, although two individuals in our cohort were siblings carrying the same disease-causing mutation in residue I871. This recurrence suggests parental mosaicism as the origin of the mutation but we were unable to detect the mutation in DNA from parental blood samples by deep sequencing. Interestingly, the first published report of SGS in 1978 described two siblings with the phenotype, which initially led to believe that this disorder was inherited autosomal recessively.[1]

SGS is a rare but clinically recognizable developmental disorder consisting of typical facial features, neurological alterations, various congenital anomalies and increased risk of malignancy. Neurological problems often encountered include severe intellectual disability, intractable epilepsy and cerebral blindness and deafness. Individuals with SGS frequently present with congenital anomalies in multiple organ systems including heart defects, kidney and/or genital malformations and bone abnormalities. Most affected individuals do not survive past childhood due to the severity of this disorder.

Four additional individuals presented a developmental phenotype with clinical characteristics overlapping classic SGS caused by atypical *SETBP1* mutations in close proximity to the canonical degron. SGS is often recognized and diagnosed based on the reminiscent clinical features and, interestingly, the amount of clinical overlap with SGS seems to be related to the position of the mutation in relation to the canonical degron. Both individuals carrying a mutation affecting SETBP1 residue S867 had the characteristic facial features of SGS, genital anomalies and seizures but no hydronephrosis. The absence of hydronephrosis in these cases is remarkable, since it is considered one of the hallmark features of SGS.[51] However, hydronephrosis was also absent in two individuals with mutations within the degron, suggesting that it should not be recognized as an obligatory feature for the diagnosis of SGS. Although both individuals with an atypical *SETBP1* mutation in residues E862 and T873 share some features with SGS, they would not have been classified as typical SGS. In both these individuals, the mutations were detected by whole exome sequencing and lead to the diagnosis of atypical SGS *a posteriori*. Therefore, as Carvalho *et al.* recently suggested,[44] previously proposed clinical diagnostic criteria for SGS may not be applicable for these cases. Due to the limited number of atypical cases, we cannot draw any conclusions on the survival and progress of the disease in these cases.

Somatic *SETBP1* mutations observed in myeloid malignancies have been shown to have a gain-of-function effect on the SETBP1 protein, leading to decreased binding of the βTrCP1 and increased protein levels.[5] This gain-of-function mechanism seems to also pertain to germline *SETBP1* mutations, as we observe that cells of individuals with germline *SETBP1* mutations have increased levels of SETBP1 protein. Furthermore, recent reports of germline chromosomal deletions and truncating mutations in *SETBP1* show that loss-of-function mutations in this gene cause a completely different phenotype from SGS.[52] Clinically, individuals with these genetic lesions present a phenotype characterized by a complete lack of expressive speech with intact receptive language abilities, decreased fine motor skills, subtle dysmorphisms and hyperactivity and autistic traits.[53]

The distribution of germline *SETBP1* mutations within the canonical degron differs from that of somatic events; mutations affecting residue I871 are significantly more frequent in the germline than somatically. Furthermore, germline *SETBP1* mutations are on average less disruptive to the βTrCP1-SETBP1 interaction than somatic mutations. Notably, the isoleucine at position 871 is a variable residue within the SETBP1 degron, which is defined as "DpSGφXpT" where φ (in this case, isoleucine) represents a hydrophobic amino acid and X stands for any residue.[18] Previous studies of the structure of the conserved destruction motif in β-catenin show that the isoleucine residue in the degron has a small role in the interaction with βTrCP1, while the first three residues of the degron and the asparagine in particular are essential for protein-protein binding.[20] Thus, the molecular differences observed between germline and somatic *SETBP1* mutations are likely caused by variation in the prevalence of mutations in residue I871 in each condition. As a consequence of the weak role of residue I871 in the interaction of SETBP1 with βTrCP1, mutations in this residue, although disruptive when present in the germline, are functionally milder.

The difference in molecular consequences between somatic and germline *SETBP1* mutations is in line with findings from previous studies examining somatic and germline mutations in *PTPN11*, involved in juvenile myelomonocytic leukemia and in Noonan syndrome, respectively.[54,55] In contrast with *SETBP1* mutations, germline *PTPN11* mutations causative for Noonan syndrome rarely overlap with somatic mutations observed in leukemia. This mutual exclusivity between germline and somatic *PTPN11* mutations has been proposed to result from the existence of distinct thresholds for gain-of-function mutations in developmental phenotypes and tumorigenesis.[55] Despite their overlap, our analysis of germline and somatic *SETBP1* mutations supports this model. Malignant cell behavior in cancer can only be driven by somatic mutations with an intense activation, while germline mutations with mild activation are sufficient to disrupt normal development. Consequently, gain-of-function mutations usually found somatically may lead to prenatal lethality or severe developmental alterations features when present in the germline, as a result of intense activation.[56] Likewise, mild hypermorphic mutations associated with

developmental disorders may be less likely to drive malignancy and are encountered less frequently as somatic events in cancer. In line with this, functional analysis of D868N, the *SETBP1* mutation most commonly found in cancer, shows that it leads to the highest increase in SETBP1 protein levels and cell proliferation, suggesting it has the strongest effect at the biochemical and cellular level. Remarkably, individuals with SGS caused by mutations in residue D868 have higher incidence of tumorigenesis with odds ratio above 9 when compared to individuals with SGS caused by other mutations. Our findings suggest that individuals with strongly activating germline mutations in *SETBP1* are at increased risk of malignancy. Due to the extremely low prevalence of SGS and the fact that malignancy is still a relatively infrequent complication of SGS, our study is limited by the small number of individuals with SGS who developed a malignancy. The correlation between strongly activating germline *SETBP1* mutations and risk of malignancy should be reproduced in a larger cohort, in order to provide accurate prognosis and personalized follow-up for individuals with germline *SETBP1* mutations.

Despite the increased risk for tumorigenesis, most individuals with SGS do not develop cancer. This observation that can be extended to individuals with developmental disorders resulting from germline mutations in genes involved in cancer when mutated somatically such as *ASXL1, EZH2* or *ARID1A.*[57–59] As mentioned previously, this observation may be in part due to the effect of germline mutations not reaching the threshold of functional activation required to drive cancer. Furthermore, early lethality could also explain why most individuals with developmental disorders caused by germline mutations in genes involved in cancer do not develop malignancies. For instance, somatic *SETBP1* mutations have been associated mainly with chronic myeloid leukemia, a disease occurring generally in individuals above the age of 60.[5,31] Individuals with SGS have a short life span, which may not allow for the accumulation of additional somatic mutations required for tumorigenesis. Additionally, the cellular context is also important for the expression of a mutation: certain cancer-driving mutations can only do so in the context of "aged" hematopoietic stem cells.[60–62] This combination of factors may explain why myeloid malignancy is a rare malignancy in SGS, while embryonic cancers are observed more frequently. Finally, the presence of mutations in all cells of the organism instead of a subset may eliminate cellular advantage in a single cell which would allow for clonal expansion and, eventually, malignancy.

Cell lines derived from individuals with SGS recapitulate some of the features identified in myeloid cells with *SETBP1* mutations, including increased SETBP1 and SET protein levels and enhanced cell proliferation.[5,7] Although we observed a *SETBP1* mutation-specific effect on cell proliferation, we did not detect visible differences in SET protein levels between individuals with different *SETBP1* mutations. It is possible that different *SETBP1* mutations have a subtle but distinct effect on SET protein levels that are not detectable by Western blot. However, we cannot rule out the absence of a mutation-specific effect of *SETBP1* mutations on

SET protein levels with additional mutation-specific downstream alterations resulting from pathways that do not involve SET protein. In light of our findings, cell lines derived from patients with developmental disorders caused by germline mutations in genes implicated in tumorigenesis may be a valuable model for the study of downstream consequences of cancer mutations at the molecular level. Furthermore, the similarities observed in the molecular consequences of germline and somatic *SETBP1* mutations support the rationale behind recent studies which have elegantly repurposed drugs used in cancer therapy for the treatment of developmental disorders in mouse models.[63] An increase in SETBP1 protein as a result of mutations or overexpression leads to an increase in SET protein which results in the inactivation of PP2A. Although currently there are no known SETBP1 antagonists, several compounds have been described to antagonize SET protein or to lead to PP2A activation.[64] For example, OP449 is a synthetic peptide that binds to SET protein, activates PP2A and selectively inhibits cell growth in leukemia cell lines and primary patient cells.[65] Other compounds that directly activate PP2A, such as FTY720, have also been described.[66] While promising, it is yet unclear whether compounds targeting SET or PP2A may prove useful in the treatment of individuals with germline SETBP1 mutations. For one, the window of time between diagnosis and potential therapeutic interventions for early developmental phenotypes may be minimal. However, although this is still a speculative point, future therapeutic interventions may be useful to prevent progression of certain features of SGS, such as neurodegeneration.[28] Finally, it is still possible that individuals with SGS present alterations in other proteins downstream of SETBP1.

In summary, we describe the largest cohort of *SETBP1* mutation-positive SGS patients to date. Our results support that typical SGS is caused by gain-of-function mutations of *SETBP1* affecting a degron involved in SETBP1 protein stability, while novel mutations outside the canonical degron cause an atypical form of SGS characterized by a milder phenotype. We observe variability in the magnitude of effect of germline mutations within the canonical degron of SETBP1 with consequences at the biochemical level and influencing the cellular and clinical phenotype. Furthermore, our results highlight that, despite the identification of overlapping *SETBP1* mutations in SGS and myeloid malignancies, the mutation spectrum is significantly different in both conditions with functionally weaker mutations appearing predominantly as germline mutations in SGS. The parallelisms between the functional consequences of germline and somatic *SETBP1* mutations is relevant for the better understanding of SGS but could also deliver insight into the role of *SETBP1* as a cancer driver. Finally, our findings highlight that the convergence of the fields of cancer and developmental disorders uncovers common molecular mechanisms of disease for overlapping germline and somatic pathogenic mutations and may support the development of drugs with a dual therapeutic role in developmental disorders and cancer.

Materials and methods

DNA studies

This study was approved by the institutional review board Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen NL36191.091.11. Written informed consent was obtained from all individuals. All legal representatives of the patients included consented to participate in our study. Genomic DNA was extracted from saliva or blood using the QIAamp DNA Mini kit (QIAGEN). The hotspot region of *SETBP1* was amplified by PCR and Sanger sequenced. Primer sequences are listed in Supplementary Table S3.

Phenotyping of individuals with SGS

Clinical features of participants were initially evaluated by clinicians from various countries. Photographs and medical information of all individuals were further assessed by a single clinical geneticist (BvB). Separate informed consent was obtained for publication of photographs.

DNA constructs

The full-length SETBP1 construct fused to a C-terminal Myc-FLAG tag was purchased from Origene (RC229443). SETBP1 variant constructs were generated using the Quikchange II Site-Directed Mutagenesis kit (Agilent). Primer sequences are listed in Supplementary table S3. SETBP1 cDNAs were subcloned using EcoRI/XhoI restriction sites into a modified pEGFP-C2 vector (Clontech) where the N-terminal EGFP tag was replaced with a YFP tag. All constructs were verified by Sanger sequencing.

Cell culture and transfection

Fibroblast cell lines were established from skin biopsies of SGS cases and controls. HEK293 cells and fibroblasts were cultured in DMEM supplemented with 10% fetal bovine serum (all from Invitrogen). LCLs were established by Epstein-Barr virus transformation of peripheral lymphocytes from blood samples of SGS patients and controls. LCLs were cultured in RPMI medium (Lonza) with 10% fetal bovine serum and 5% HEPES. Transfections were performed using the GeneJuice transfection reagent following the manufacturer's instructions (Merck Millipore).

Western blotting

Whole-cell lysates were prepared as described previously [67]. Total protein was quantified using the Pierce™ BCA protein assay kit (ThermoFisher Scientific). Proteins were resolved on 4-15% Tris-Glycine gels and transferred to PDVF membranes (Bio-Rad). After blotting, membranes were incubated overnight at 4^0C with the appropriate primary antibodies: rabbit anti-SETBP1 (Santa Cruz, sc-85148, 1:100), rabbit anti-SET (Abcam, ab1183, 1:4000), mouse anti-β-actin (Sigma, AC-15, 1:10000) and mouse anti-FLAG (Sigma, F1804; 1:1000). Membranes

were then incubated with HRP-conjugated donkey anti-rabbit (Abcam) or goat anti-mouse (Bio-Rad) secondary antibodies. Proteins were visualized using the Novex ECL Chemiluminescent Substrate Reagent kit (Invitrogen) and the ChemiDoc XRS+ System (Bio-Rad).

Protein stability assays

Cells were transfected in clear-bottomed black 96-well plates in triplicate with YFP-SETBP1 expression plasmids together with a modified pmCherry-C1 plasmid to normalize for transfection efficiency. Fourteen hours post-transfection, YFP and mCherry fluorescence intensities were measured for 24h in live cells in a TECAN M200PRO microplate reader at 37°C and 5% CO_2. In the case of stability assays with proteasome inhibitor, MG132 (Sigma) was added to the culture medium (10uM final concentration) 48 hours post-transfection and fluorescence intensities were measured at 0, 3 and 12 hours.

Fluorescence microscopy

Cells were grown on poly-L-lysine (Sigma) coated coverslips and were fixed 48 hours post-transfection with 4% paraformaldehyde (Electron Microscopy Sciences) for 10 minutes at room temperature. YFP was visualized by direct fluorescence. Nuclei were visualized with DAPI (Vectorlabs). Fluorescence images were obtained using an Axio Imager Z1 fluorescence microscope (Zeiss).

Proliferation assays

Lymphoblastoid cell lines were synchronized by overnight serum starvation, after which they were seeded in 24-well plates at a concentration of 160,000 cells/mL (three replicates per cell line). A measurement was performed every 24 hours, in which the cells were mixed with Trypan blue and counted.

Modeling of protein in Yasara and FoldX

The SETBP1 degron variants observed in SGS and in leukemia were manually curated from previously published reports) [4–7,23–27,30–48]. Mutations were modeled using the YASARA structural simulation software (http://www.yasara.org/). The protein model for βTrCP1 and degron sequence from β-catenin was obtained from the RCSB protein bank (1p22) [18]. The FoldX plugin for YASARA was used to calculate ΔΔG values [68].

RNA-sequencing

LCLs from individuals with SGS were derived from blood samples of individuals with germline *SETBP1* mutations D868N or I871T (n=2 for each). LCLs from controls (n=8) were derived from blood samples of individuals with different forms of intellectual disability [69]. RNA was purified from cell cultures using RNeasy mini kit from QIAGEN and RNAseq libraries were prepared using the TruSeq Stranded mRNA kit (Illumina) and sequenced on a Nextseq platform using

2x75 bp paired end sequencing. RNA-seq samples were mapped using Kallisto (version 0.42.4) to the human genome (hg19.75). At least 70% of the reads for each sample were mapped to the human genome entailing at least 39.6 million reads in the sample with the lowest sequencing depth. Differential gene expression was performed using DESeq2 [49]. Gene set enrichment analysis was performed using the GSEA software [50].

Statistics

Statistical analysis was performed using R Statistical Software (http://www.r-project.org/ version 3.1.2, R Foundation for Statistical Computing, Vienna, Austria). Student's T test was used for comparison of two groups with normal distribution, otherwise Mann-Whitney's U test was used. For comparison of multiple groups against each other, we used ANOVA and Tukey's test. To determine whether the distribution of mutations within the degron was the same in germline and somatic mutations, we used Fisher's test and performed Bonferroni multiple test correction.

References

1. Schinzel, A. & Giedion, A. A syndrome of severe midface retraction, multiple skull anomalies, clubfeet, and cardiac and renal malformations in sibs. *Am J Med Genet* **1,** 361–375 (1978).
2. Minn, D. *et al.* Further clinical and sensorial delineation of Schinzel-Giedion syndrome: Report of two cases. *Am J Med Genet* **109,** 211–217 (2002).
3. Al-Mudaffer, M. *et al.* Clinical and radiological findings in Schinzel-Giedion syndrome. *Eur J Pediatr* **167,** 1399–1407 (2008).
4. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* **42,** 483–485 (2010).
5. Piazza, R. *et al.* Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat Genet* **45,** 18–24 (2012).
6. Sakaguchi, H. *et al.* Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. *Nat Genet* **45,** 937–41 (2013).
7. Makishima, H. *et al.* Somatic SETBP1 mutations in myeloid malignancies. *Nat Genet* **45,** 942–6 (2013).
8. Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat Neurosci* **17,** 764–772 (2014).
9. Agha, M. M. *et al.* Congenital abnormalities and childhood cancer. *Cancer* **103,** 1939–1948 (2005).
10. Bjorge, T., Cnattingius, S., Lie, R. T., Tretli, S. & Engeland, A. Cancer Risk in Children with Birth Defects and in Their Families: A Population Based Cohort Study of 5.2 Million Children from Norway and Sweden. *Cancer Epidemiol Biomarkers Prev* **17,** 500–506 (2008).
11. Merks, J. H. M. *et al.* Prevalence and Patterns of Morphological Abnormalities in Patients With Childhood Cancer. *JAMA* **299,** 61–69 (2008).
12. Durmaz, A. *et al.* The Association of minor congenital anomalies and childhood cancer. *Pediatr Blood Cancer* **56,** 1098–102 (2011).
13. Kelleher, F. C., Fennelly, D. & Rafferty, M. Common critical pathways in embryogenesis and cancer. *Acta Oncol (Madr)* **45,** 375–388 (2006).
14. Ravid, T. & Hochstrasser, M. Diversity of degradation signals in the ubiquitin–proteasome system. *Nat Rev Mol Cell Biol* **9,** 679–689 (2008).
15. Inoue, D. *et al.* SETBP1 mutations drive leukemic transformation in ASXL1-mutated MDS. *Leukemia* **29,** 847–857 (2015).
16. Cristobal, I. *et al.* SETBP1 overexpression is a novel leukemogenic mechanism that predicts adverse outcome in elderly patients with acute myeloid leukemia. *Blood* **115,** 615–625 (2010).
17. Oakley, K. *et al.* Setbp1 promotes the self-renewal of murine myeloid progenitors via activation of Hoxa9 and Hoxa10. *Blood* **119,** 6099–6108 (2012).
18. Wu, G. *et al.* Structure of a beta-TrCP1-Skp1-beta-catenin complex: destruction motif binding and lysine specificity of the SCF(beta-TrCP1) ubiquitin ligase. *Mol Cell* **11,** 1445–56 (2003).
19. Verloes, A. *et al.* Schinzel-Giedion syndrome. *Eur J Pediatr* **152,** 421–3 (1993).
20. Kondoh, T. *et al.* A case of Schinzel-Giedion syndrome complicated with progressive severe gingival hyperplasia and progressive brain atrophy. *Pediatr Int* **43,** 181–184 (2001).
21. Carvalho, E. *et al.* Schinzel-Giedion syndrome in two Brazilian patients: Report of a novel mutation in SETBP1 and literature review of the clinical features. *Am J Med Genet Part A* **167,** 1039–1046 (2015).
22. Suphapeetiporn, K., Srichomthong, C. & Shotelersuk, V. SETBP1 mutations in two Thai patients with Schinzel-Giedion syndrome. *Clin Genet* **79,** 391–393 (2011).
23. Lestner, J. M., Chong, W. K., Offiiah, A., Kefas, J. & Vandersteen, A. M. Unusual neuroradiological features in Schinzel-Giedion syndrome: a novel case. *Clin Dysmorphol* **21,** 152–4 (2012).
24. Ko, J. M. *et al.* Distinct neurological features in a patient with Schinzel–Giedion syndrome caused by a recurrent SETBP1 mutation. *Child's Nerv Syst* **29,** 525–529 (2013).
25. Miyake, F. *et al.* West Syndrome in a Patient With Schinzel-Giedion Syndrome. *J Child Neurol* **30,** 932–936 (2015).
26. López-González, V. *et al.* [Schinzel-Giedion syndrome: a new mutation in SETBP1]. *An Pediatr*

(Barc) **82,** e12-6 (2015).
27. Volk, A., Conboy, E., Wical, B., Patterson, M. & Kirmani, S. Whole-Exome Sequencing in the Clinic: Lessons from Six Consecutive Cases from the Clinician's Perspective. *Mol Syndromol* **6,** 23–31 (2015).
28. Takeuchi, A. *et al.* Progressive brain atrophy in Schinzel–Giedion syndrome with a SETBP1 mutation. *Eur J Med Genet* **58,** 369–371 (2015).
29. Herenger, Y. *et al.* Long term follow up of two independent patients with Schinzel–Giedion carrying SETBP1 mutations. *Eur J Med Genet* **58,** 479–487 (2015).
30. Damm, F. *et al.* SETBP1 mutations in 658 patients with myelodysplastic syndromes, chronic myelomonocytic leukemia and secondary acute myeloid leukemias. *Leukemia* **27,** 1401–3 (2013).
31. Laborde, R. R. *et al.* SETBP1 mutations in 415 patients with primary myelofibrosis or chronic myelomonocytic leukemia: independent prognostic impact in CMML. *Leukemia* **27,** 2100–2 (2013).
32. Pardanani, a *et al.* CSF3R T618I is a highly prevalent and specific mutation in chronic neutrophilic leukemia. *Leukemia* **27,** 1870–1873 (2013).
33. Meggendorfer, M. *et al.* SETBP1 mutations occur in 9% of MDS/MPN and in 4% of MPN cases and are strongly associated with atypical CML, monosomy 7, isochromosome i(17)(q10), ASXL1 and CBL mutations. *Leukemia* **27,** 1852–1860 (2013).
34. Thol, F. *et al.* SETBP1 mutation analysis in 944 patients with MDS and AML. *Leukemia* **27,** 2072–2075 (2013).
35. Fernandez-Mercado, M. *et al.* Mutations in SETBP1 are recurrent in myelodysplastic syndromes and often coexist with cytogenetic markers associated with disease progression. *Br J Haematol* **163,** 235–9 (2013).
36. Shiba, N. *et al.* SETBP1 mutations in juvenile myelomonocytic leukaemia and myelodysplastic syndrome but not in paediatric acute myeloid leukaemia. *Br J Haematol* **164,** 156–159 (2014).
37. Hou, H.-A. A. *et al.* Clinical implications of the SETBP1 mutation in patients with primary myelodysplastic syndrome and its stability during disease progression. *Am J Hematol* **89,** 181–6 (2014).
38. Lasho, T. L. *et al.* Chronic neutrophilic leukemia with concurrent CSF3R and SETBP1 mutations: single colony clonality studies, in vitro sensitivity to JAK inhibitors and lack of treatment response to ruxolitinib. *Leukemia* **1,** 1–3 (2014).
39. Senín, A. *et al.* [Molecular characterization of atypical chronic myeloid leukemia and chronic neutrophilic leukemia]. *Med Clin (Barc)* **144,** 487–90 (2015).
40. Fabiani, E. *et al.* SETBP1 mutations in 106 patients with therapy-related myeloid neoplasms. *Haematologica* **99,** e152–e153 (2014).
41. Ammatuna, E. *et al.* Atypical chronic myeloid leukemia with concomitant CSF3R T618I and SETBP1 mutations unresponsive to the JAK inhibitor ruxolitinib. *Ann Hematol* **94,** 879–880 (2015).
42. Elliott, M. A. *et al.* ASXL1 mutations are frequent and prognostically detrimental in CSF3R-mutated chronic neutrophilic leukemia. *Am J Hematol* **90,** 653–656 (2015).
43. Cui, Y. *et al.* CSF3R, SETBP1 and CALR mutations in chronic neutrophilic leukemia. *J Hematol Oncol* **7,** 77 (2014).
44. Gambacorti-Passerini, C. B. *et al.* Recurrent ETNK1 mutations in atypical chronic myeloid leukemia. *Blood* **125,** 499–503 (2015).
45. Bartels, S. *et al.* De novo CSF3R mutation associated with transformation of myeloproliferative neoplasm to atypical CML. *Ann Hematol* **94,** 1255–1256 (2015).
46. Maxson, J. E. *et al.* The Colony-Stimulating Factor 3 Receptor T640N Mutation Is Oncogenic, Sensitive to JAK Inhibition, and Mimics T618I. *Clin Cancer Res* **22,** 757–764 (2016).
47. Lasho, T. L., Elliott, M. A., Pardanani, A. & Tefferi, A. CALR mutation studies in chronic neutrophilic leukemia. *Am J Hematol* **89,** 450 (2014).
48. Xu, L. *et al.* Genomic landscape of CD34+ hematopoietic cells in myelodysplastic syndrome and gene mutation profiles as prognostic markers. *Proc Natl Acad Sci U S A* **111,** 8589–94 (2014).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).

50. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* (2005). doi:10.1073/pnas.0506580102
51. Lehman, A. M. *et al.* Schinzel-Giedion syndrome: Report of splenopancreatic fusion and proposed diagnostic criteria. *Am J Med Genet Part A* **146,** 1299–1306 (2008).
52. Filges, I. *et al.* Reduced expression by SETBP1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome. *J Med Genet* **48,** 117–122 (2011).
53. Barnett, C. P. & van Bon, B. W. M. Monogenic and chromosomal causes of isolated speech and language impairment. *J Med Genet* **52,** 719–29 (2015).
54. Kratz, C. P. *et al.* The mutational spectrum of PTPN11 in juvenile myelomonocytic leukemia and Noonan syndrome/myeloproliferative disease. *Blood* **106,** 2183–5 (2005).
55. Tartaglia, M. *et al.* Diversity and functional consequences of germline and somatic PTPN11 mutations in human disease. *Am J Hum Genet* **78,** 279–90 (2006).
56. Schubbert, S., Shannon, K. & Bollag, G. Hyperactive Ras in developmental disorders and cancer. *Nat Rev Cancer* (2007). doi:10.1038/nrc2109
57. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* **43,** 729–731 (2011).
58. Tsurusaki, Y. *et al.* Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nat Genet* **44,** 376–378 (2012).
59. Gibson, W. T. *et al.* Mutations in EZH2 Cause Weaver Syndrome. *Am J Hum Genet* **90,** 110–118 (2012).
60. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N Engl J Med* **371,** 2477–2487 (2014).
61. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep* **10,** 1239–1245 (2015).
62. Mason, C. C. *et al.* Age-related mutations and chronic myelomonocytic leukemia. *Leukemia* (2015). doi:10.1038/leu.2015.337
63. Bjornsson, H. T. *et al.* Histone deacetylase inhibition rescues structural and functional brain deficits in a mouse model of Kabuki syndrome. *Sci Transl Med* **6,** 256ra135-256ra135 (2014).
64. Neviani, P. & Perrotti, D. SETting OP449 into the PP2A-activating drug family. *Clin Cancer Res* **20,** 2026–2028 (2014).
65. Meckenzie, R. J. *et al.* OP449, a Novel SET Antagonist, Is Cytotoxic To Leukemia Cells and Enhances Efficacy Of Tyrosine Kinase Inhibitors In Drug-Resistant Myeloid Leukemias. *Blood* **122,** 2511 LP-2511 (2013).
66. Perrotti, D. & Neviani, P. Protein phosphatase 2A: A target for anticancer therapy. *Lancet Oncol* **14,** e229–e238 (2013).
67. Deriziotis, P. *et al.* De novo TBR1 mutations in sporadic autism disrupt protein functions. *Nat Commun* **5,** 4954 (2014).
68. Van Durme, J. *et al.* A graphical interface for the FoldX forcefield. *Bioinformatics* **27,** 1711–1712 (2011).
69. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).

Supplementary data

Supplementary tables

Gene name	Uniprot accession number	Sequence of βTrCP binding site (DpSGφXpS/pT)	Motif start position	Motif stop position
SETBP	Q9Y6X0	DSGIGT	868	873
CTNNB1	P35222	DSGIHS	32	37
NFE2L2	Q16236	DSGISL	343	348
ATF4	P18848	DSGICM	218	223
EEF2K	O00418	DSGYPS	440	445
CLSPN	Q9HAW4	DSGQGS	29	34
REST	Q13127	DEGIHS	1008	1013
CDC25A	P30304	DSGFCL	81	86
PDCD4	Q53EL6	DSGRGD	70	75
PER1	O15534	TSGCSS	121	126

Supplementary Table S1. Protein sequence alignment of degrons targeted by βTrCP. Based on Low TY, Peng M, Magliozzi R, Mohammed S, Guardavaccaro D, Heck AJR. A systems-wide screen identifies substrates of the SCF bTrCP ubiquitin ligase. 2014;7: 1–12.

Mutation	Stability ΔΔG (kcal/mol)	ST DEV	βTrCP1 Interaction ΔΔG (kcal/mol)	ST DEV	Interpretation	Germline	Somatic
D868A	1.08	0.24	2.78	0.14	Destabilizing	Y	N
D868G	0.99	0.21	2.72	0.21	Destabilizing	N	Y
D868H	9.98	3.04	10.77	3.43	Highly destabilizing	N	Y
D868N	2.30	0.17	3.57	0.49	Highly destabilizing	Y	Y
D868Y	12.31	2.88	11.74	2.77	Highly destabilizing	Y	Y
S869G	0.10	0.11	0.06	0.11	Neutral	N	Y
S869N	-0.25	0.12	-0.47	0.21	Neutral	Y	Y
S869R	-0.97	0.16	-1.05	0.17	Neutral	Y	Y
G870C	1.27	1.04	0.79	0.98	Destabilizing	Y	N
G870D	4.94	1.46	4.78	1.48	Highly destabilizing	Y	Y
G870R	11.72	1.81	10.84	2.02	Highly destabilizing	N	Y
G870S	1.07	0.51	0.65	0.50	Destabilizing	Y	Y
G870V	5.40	2.12	5.21	2.05	Highly destabilizing	N	Y
I871S	0.17	0.31	-0.38	0.34	Neutral	Y	Y
I871T	-0.22	0.42	-1.15	0.34	Neutral	Y	Y

Supplementary Table S2. ΔΔG values for protein stability and degron-βTrCP1 interaction for SETBP1 variants reported in SGS cases and in myeloproliferative disorders. ST DEV: Standard deviation.

Name	Sequence (5'-3')	Purpose
SETBP1 Exon4 F	CTTCACCAGCAGCTATGCAC	Sanger sequencing
SETBP1 Exon4 R	CGGTGGGAGATTCTGAACAC	Sanger sequencing
SETBP1 E862K F	GAGTCCCACAGTAAGGAGACGATCCC	Site-directed mutagenesis
SETBP1 E862K R	GGGGATCGTCTCCTTACTGTGGGACTC	Site-directed mutagenesis
SETBP1 D868N F	GACGATCCCCAGCAACAGCGGCATTGG	Site-directed mutagenesis
SETBP1 D868N R	CCAATGCCGCTGTTGCTGGGGATCGTC	Site-directed mutagenesis
SETBP1 S869N F	GACGATCCCCAGCGACAACGGCATTGG	Site-directed mutagenesis
SETBP1 S869N R	CCAATGCCGTTGTCGCTGGGGATCGTC	Site-directed mutagenesis
SETBP1 G870S F	AGCGACAGCAGCATTGGGACAGAC	Site-directed mutagenesis
SETBP1 G870S R	GTCTGTCCCAATGCTGCTGTCGCT	Site-directed mutagenesis
SETBP1 I871T F	GACAGCGGCACTGGGACAGACAAC	Site-directed mutagenesis
SETBP1 I871T R	GTTGTCTGTCCCAGTGCCGCTGTC	Site-directed mutagenesis

Supplementary Table S3. Sequences of primers used in this study.

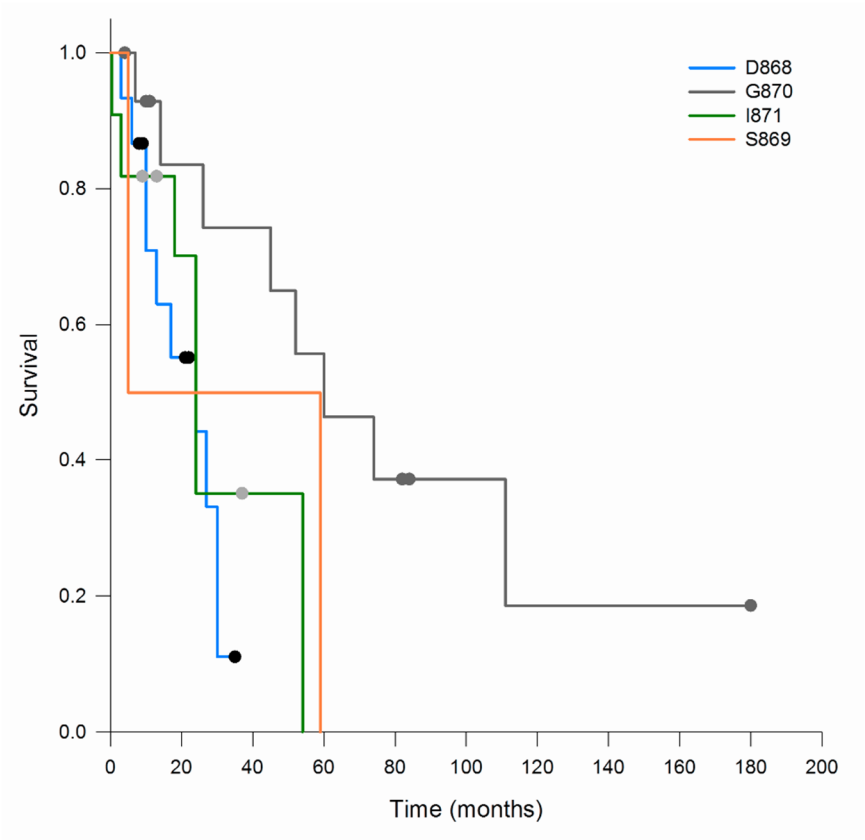
Name	Size	ES	NES	FDR q-val	FWER p-val
Reactome peptide chain elongation	100	0.54	6.55	<0.00001	<0.001
Reactome influenza viral RNA transcription and replication	116	0.47	6.08	<0.00001	<0.001
Reactome 3 UTR mediated translational regulation	121	0.46	5.82	<0.00001	<0.001
Reactome nonsense mediated decay enhanced by the exon junction complex	121	0.44	5.62	<0.00001	<0.001
Reactome translation	163	0.35	5.44	<0.00001	<0.001
Reactome SRP dependent cotranslational protein targeting to membrane	124	0.42	5.32	<0.00001	<0.001
Reactome influenza life cycle	150	0.35	4.95	<0.00001	<0.001
Reactome metabolism of proteins	427	0.16	3.95	<0.00001	<0.001
Reactome formation of the ternary complex and subsequently the 43s complex	53	0.44	3.80	<0.00001	<0.001
Reactome activation of the mRNA upon binding of the cap binding complex	61	0.40	3.75	<0.00001	<0.001
Reactome metabolism of mRNA	224	0.20	3.49	<0.00001	<0.001
Reactome respiratory electron transport ATP synthesis by chemiosmotic coupling and heat production by uncoupling proteins	80	0.33	3.42	<0.00001	<0.001
Reactome TCA cycle and respiratory electron transport	116	0.27	3.40	<0.00001	<0.001
Reactome cholesterol biosynthesis	21	0.62	3.39	<0.00001	<0.001
Reactome metabolism of RNA	269	0.17	3.30	<0.00001	<0.001
Reactome respiratory electron transport	65	0.33	3.18	<0.00001	<0.001
Reactome telomere maintenance	73	0.32	3.15	<0.00001	<0.001
Reactome generic transcription pathway	323	0.14	2.94	1.28E-04	0.002
Reactome cell cycle	396	0.13	2.90	1.22E-04	0.002
Reactome glycolysis	25	0.48	2.89	1.15E-04	0.002
Reactome chromosome maintenance	116	0.23	2.84	2.28E-04	0.004
Reactome RNA Pol I promoter opening	59	0.30	2.75	3.28E-04	0.006
Reactome gluconeogenesis	29	0.41	2.63	7.19E-04	0.014
Reactome meiotic recombination	80	0.25	2.59	0.001	0.024
Reactome packaging of telomere ends	46	0.32	2.53	0.001	0.039

Supplementary Table S4. Gene set enrichment analysis for genes differentially expressed in LCLs from individuals with germline *SETBP1* mutations D868N or I871T and LCLs from controls. Gene set enrichment analysis was performed as described in Subramanian et al. *PNAS* 2005 using curated gene sets from the reactome pathway database. An enrichment is observed for genes involved in mRNA transcription and translation, mitochondrial respiration and in the cell cycle. While no enrichment was observed for specific pathways, a closer examination of the genes involved in cell cycle that we detected to be enriched in the samples of individuals with germline *SETBP1* mutations revealed that the genes identified were associated with mitosis, sister chromatid cohesion, DNA replication and nucleosome assembly.

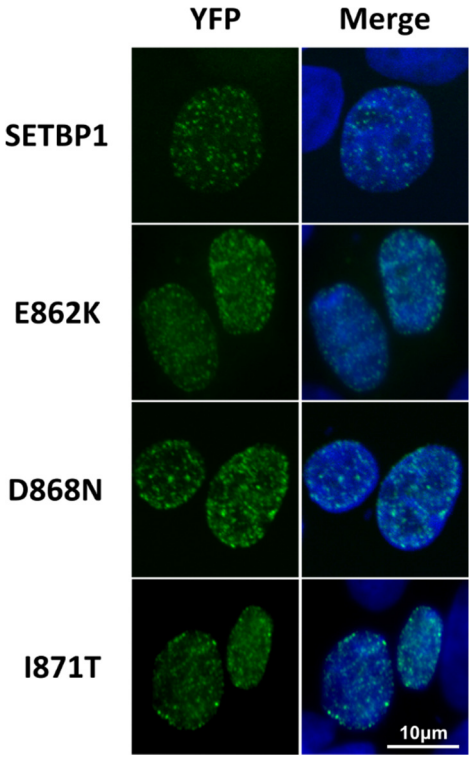
Supplementary figures

Supplementary Figure S1. Facial features in typical Schinzel-Giedion syndrome. Panels A to I show the facial features of patients with germline *SETBP1* mutations within the canonical degron leading to typical SGS (A: current case 1; B: current case 2; C: current case 3; D: current case 12; E: current case 8; F: current case 16; G: current case 18; H current case 20; I current case 23). Progression of facial features with age in current case 9 with germline SETBP1 mutation G870S at 1 week of age (J), at 1 ½ years of age (K), at 6 years of age (L) and at 7 years of age (M).

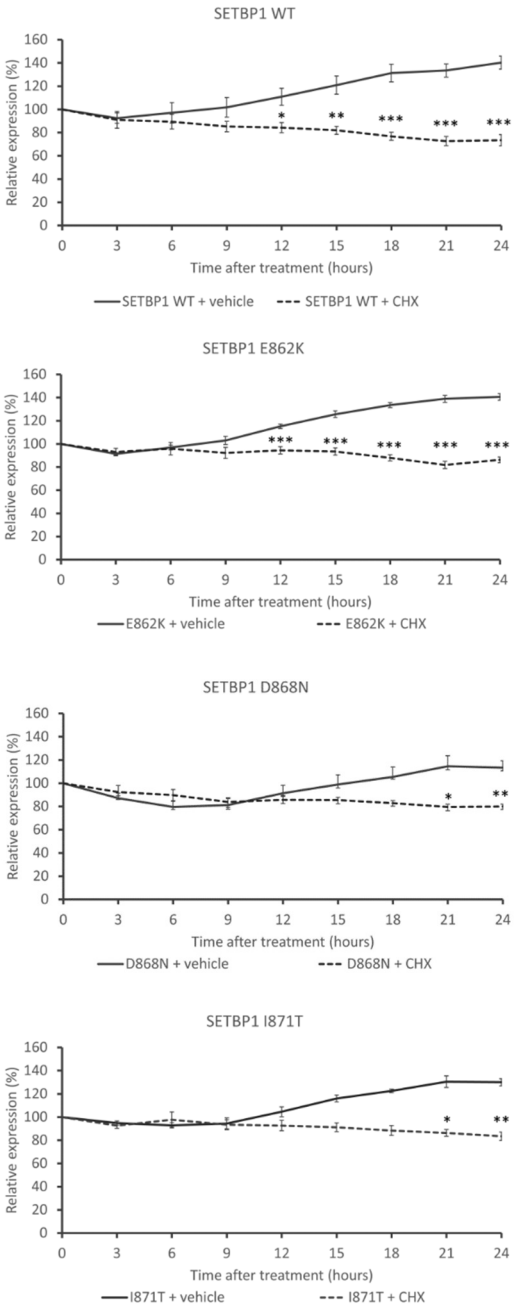
Supplementary Figure S2. Phenotype of individuals with germline *SETBP1* mutations outside the canonical degron leading to atypical Schinzel-Giedion. Individual with germline SETBP1 mutation S867R (current case 27) at 1 week of age (A), at nine months (B), at two years of age (C) and at four years of age (D). Patient with germline SETBP1 mutation E862K (current case 28) at 5 years of age (E and F). Right hand and foot (G and H), note the small feet with short toes. Individual with germline SETBP1 mutation T873I (current case 29) at one month of age (I), at eleven months (J), at three years of age (K) and at four years of age (L).



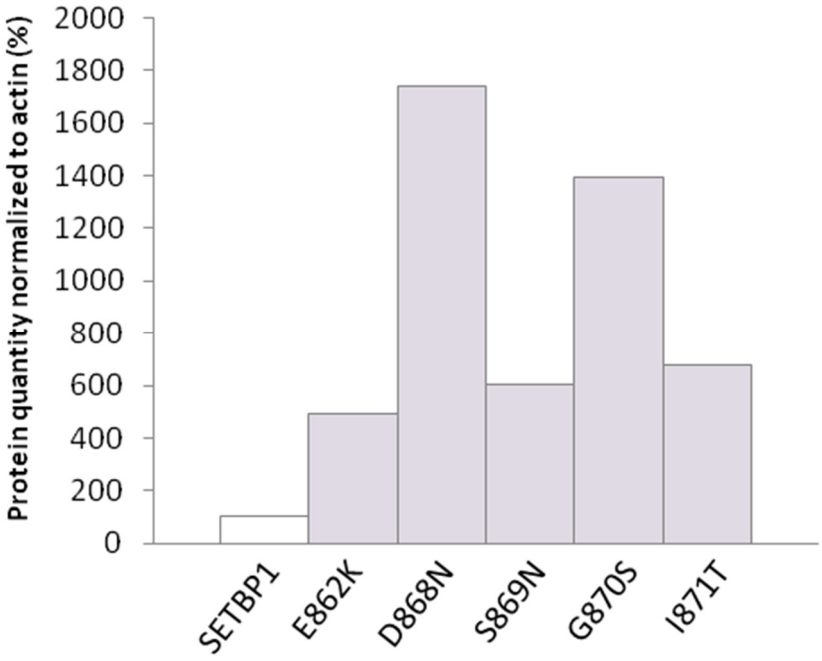
Supplementary Figure S3. LogRank survival analysis performed for 44 individuals with SGS for which data was available from our own cohort and previous reports. Dots represent censored data, for age at which the individual was last known to be alive. The group of patients with mutations in residue G870 had a statistically significantly longer survival than individuals with mutations in residue D868 (median of 60 months versus 24 months respectively, $p<0.01$).



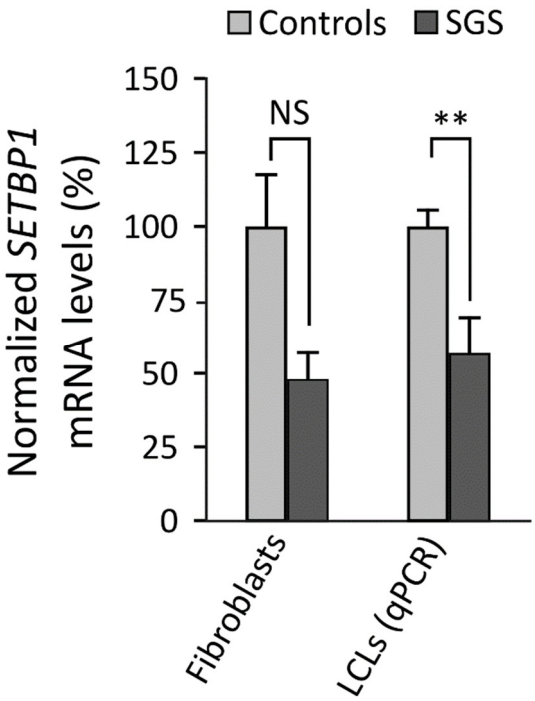
Supplementary Figure S4. Fluorescence imaging of HEK293 cells expressing YFP-tagged SETBP1 variants (green). Nuclei were stained with DAPI (blue). Wild-type (WT) SETBP1 and disease-causing SETBP1 variants were expressed in HEK293 cells as YFP fusion proteins. Direct fluorescence imaging of SETBP1 variants showed that the WT protein localizes to the nucleus, with a speckle-like pattern typical of chromatin-interacting proteins. Pathogenic SETBP1 protein variants occurring within (D868N, I871) or in close proximity (E862K) to the canonical degron sequence had no effect on protein localization.



Supplementary Figure S5. Relative expression of SETBP1 protein variants in live HEK293 cells treated with 50ug/uL cycloheximide or vehicle only as controls. Bars represent the standard error. (* $p<0.05$, ** $p<0.01$, *** $p<0.001$ versus cells at same time point treated with vehicle; Student's T test and Mann-Whitney U test for measurements with non-normal distribution).

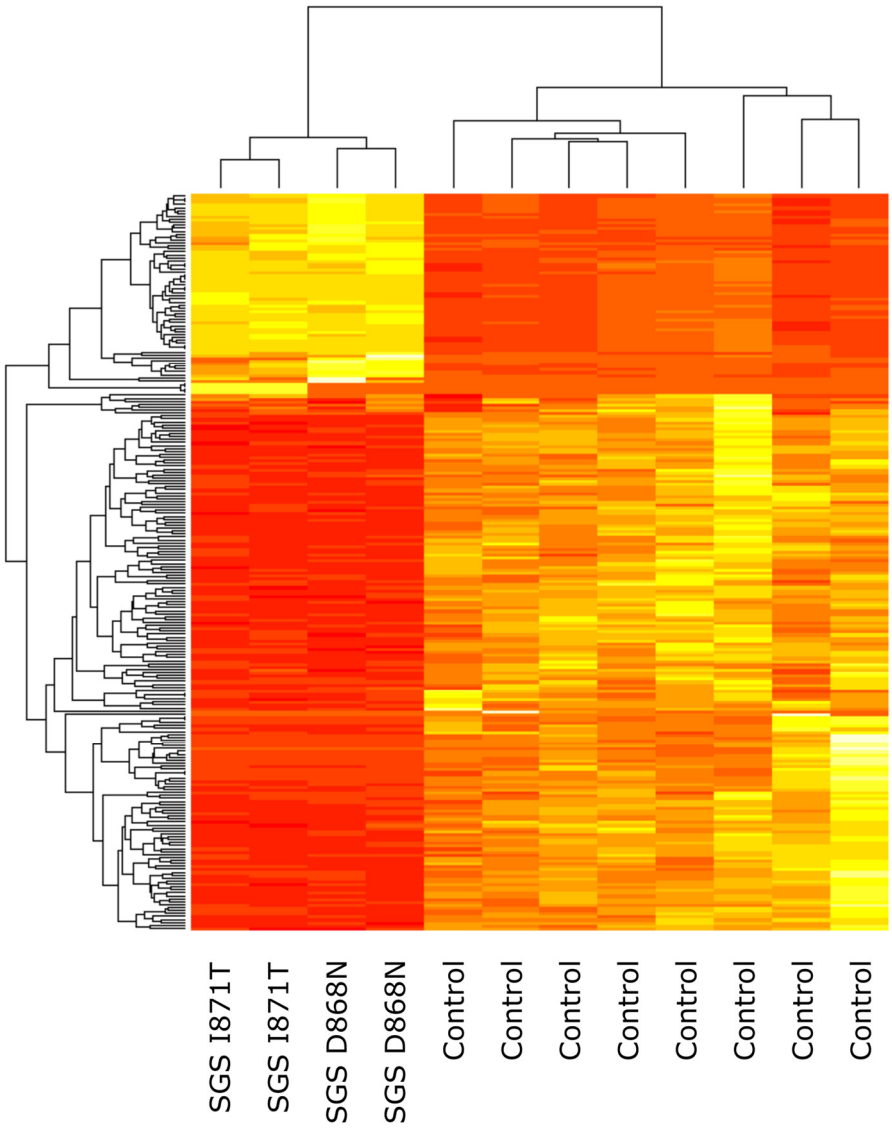


Supplementary Figure S6. Densitometry for Western blot of overexpressed wild-type and mutant SETBP1-FLAG in HEK293 cells using actin for normalization.



Supplementary Figure S7. Levels of *SETBP1* mRNA in fibroblasts and lymphoblastoid cell lines (LCLs) from individuals with SGS as compared to controls. *SETBP1* mRNA levels were normalized to *ACTB* and *GAPDH*. The results shown represent the mRNA levels in fibroblasts cells lines from 2 controls, fibroblast cell lines from 2 individuals with SGS, LCLs from 2 controls and LCLs from 3 individuals with SGS. The bars represent the standard error. NS: not significant. ** $p<0.01$, Student's T test.

The decrease in *SETBP1* mRNA levels observed in individuals with SGS compared to controls could be the result of a feedback mechanism in which increased SETBP1 protein levels may lead to a reduction in *SETBP1* transcription thereby decreasing SETBP1 protein levels. A similar observation was reported for myeloid progenitor cells immortalized by mutant *SETBP1* compared to cells immortalized by wild-type *SETBP1* in Makishima H. *et al.* Nat Genetics 2013. We speculate that this mechanism would be able to lower SETBP1 protein to normal levels for SETBP1 harboring weaker mutations, such as I871T, but not for stronger mutations, such as D868N.



Supplementary Figure S8. Heatmap of RNAseq transcriptome analysis showing differential gene expression between LCLs from controls (n=8) and LCLs from individuals with germline *SETBP1* mutations D868N or I871T (n=2 for each). Only the top 250 most significant genes are shown in this figure. Yellow indicates high expression while red indicates low expression. Differential gene expression was performed using DESEq2, which showed differential expression of 1,811 genes (adjusted p-value <0.01), of which 632 are upregulated and 1,179 are downregulated

Human brain. The insula of the left side is exposed by removing the opercula.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 7:
General discussion

Next generation sequencing, a revolution in human genetics

Over the last decade, the field of human genetics has enthusiastically embraced Next Generation Sequencing (NGS) both in the clinic and in research. Whole exome sequencing (WES) was first used for clinical diagnosis of a patient in 2009,[1] followed shortly after by a report in which the application of WES led for the first time to the identification of genes responsible for dominant and recessive monogenic diseases.[2,3] Since then, the implementation of NGS and further technical developments have been instrumental to identify mutations linked to hundreds of human genetic diseases, including autosomal dominant, recessive and X-linked disorders.[4,5] It is currently used routinely in medical centers around the world for the diagnosis of human genetic diseases, including pediatric disorders, neurodevelopmental and neurological diseases, sensory disorders and familial predisposition to cancer. NGS of cell-free DNA in blood can be used for non-invasive prenatal diagnosis,[6] as a promising replacement for chorio- and amniocentesis, but also to screen for cancer by searching for tumor DNA in blood as part of a so-called liquid biopsy.[7] Furthermore, sequencing of a tumor offers the possibility of analyzing its genetic characteristics to help guide treatment.[8] Finally, NGS is a powerful tool that, by studying the transcriptome,[9] the epigenome[10] and even chromatin interactions,[11] can help understand the consequences of genetic mutations in human biology and disease. All in all, the application of NGS in medicine opens the door for personalized medical interventions while shedding light on the biology of human diseases.

While the development of NGS has made it possible to sequence entire genomes with relative ease, we are only beginning to interpret the effect of genetic variation on phenotypes. The pathogenicity of a mutation is determined by different factors, such as the type of mutation, the genomic region in which it arises and the context in which it occurs, including the genetic background and the timing. In this thesis, I have made use of NGS to identify mutations leading to human disease, focusing on *de novo* mutations, the time at which they arise throughout life and the influence of timing in the consequences of mutations in human health and disease.

De novo mutations in human biology and human disease

NGS uncovers the role of de novo mutations in disease

NGS-based studies have revealed that *de novo* mutations arise in each generation and have established that pathogenic mutations with *de novo* occurrence represent a frequent cause of disease for both common and rare disorders with sporadic appearance. I discuss the different types of human disease that can be caused by *de novo* mutations in **Chapter 2**, including pediatric and adult-onset diseases. However, phenotypes associated with strongly reduced fitness and reproductive lethality, such as severe and early-onset disorders, are the human diseases most prominently associated with *de novo* mutations.

A large number of genes in which *de novo* mutations cause rare and sporadic developmental disorders have been identified through NGS, guided either by a genotype-first or a phenotype-first approach. The former relies on trio-based WES and whole genome sequencing (WGS) of an individual and both of his or her parents which can be used to identify *de novo* germline mutations throughout the exome or genome.[3] The latter is based on the use of WES and WGS to identify mutations shared by two or more unrelated individuals with the same phenotype or disorder. In **Chapter 3**, we followed a phenotype-first approach and performed WES in two unrelated individuals sharing the same clinical phenotype. We identified truncating mutations in the gene *THRA* in both individuals, which were later shown to be *de novo* by Sanger sequencing. These patients present with growth retardation, characteristic facial features, mild intellectual disability, constipation and skeletal alterations which together form a clinically well-defined thyroid hormone resistance syndrome. This study shows that a phenotype-first approach can be used for disease-gene identification but it requires great precision in the clinical phenotyping and selection of patients for WES or WGS. In our study, the two patients with the most severe phenotype were selected for WES, which resulted in the identification of truncating mutations in the same gene. However, three additional individuals presented the same clinical phenotype, among which two were found to have a missense mutation in *THRA* and one did not present mutations in *THRA*. Thus, from the point of view of disease, clinical homogeneity increases the likelihood of identifying rare mutations in the same gene in more than one individual. Furthermore, individuals with the same clinical phenotype can share mutations in two or more genes, thus requiring additional analysis to determine the most likely candidate for pathogenicity.

Most of the issues pertaining to a phenotype-first approach can be circumvented by trio-based sequencing. For one, trio-based analysis allows the identification of candidate pathogenic mutations in single cases presenting a highly heterogeneous phenotype. If *de novo* mutations are expected to be the likely cause of a disease, trio-based sequencing significantly diminishes the

number of candidate variants from more than four million to less than one hundred mutations per genome and 1 to 2 per exome.[12–16] However, as *de novo* mutations arise in all individuals and most of them do not result in disease, the identification of *de novo* mutations is not sufficient by itself to establish disease causality. Therefore, *de novo* mutations identified by trio-based sequencing need to be interpreted bioinformatically, functionally, clinically and statistically.[17] For instance, *de novo* mutations can be evaluated by *in silico* prediction programs like SIFT, PolyPhen, MutationTaster and CADD.[17–20] Additionally, bioinformatic approaches leveraging large scale databases of genetic variation can be used to assist in the interpretation of *de novo* mutations.[21,22] Pathogenic *de novo* mutations involved in early-onset severe disorders are unlikely to be found in population databases, such as the Exome Aggregation Consortium (ExAC).[23] Furthermore, constraint scores established from the population frequency of mutations in each gene can be used to indirectly estimate the pathogenicity of a mutation in a gene.[22] Functional validation to link a gene or a mutation to a phenotype can provide robust experimental evidence of the pathogenicity of a mutation, but it is often laborious and the necessary assays may differ per gene and per mutation. Many recent scientific developments can assist in the interpretation of *de novo* mutations in human disease. For instance, experimental studies can be performed on induced pluripotent stem cells from patient-derived samples which can be differentiated into cells and tissues relevant for the specific disease.[24] Additionally, CRISPR/Cas9-based methods for *in vitro* and *in vivo* genetic manipulation offer great promise in the experimental validation of mutations.[25,26] From the clinical point of view, the phenotype of patients in whom we identify candidate *de novo* mutations needs to be compared to that of other individuals with similar mutations in the same genomic region, either within the study cohort or among those previously reported in literature.[27,28] This phenotypic comparison is vital to determine whether individuals with similar mutations have overlapping phenotypic features underlying the same disorder. Finally, statistical analysis should be performed to determine whether the number of *de novo* mutations identified within a gene in a study cohort exceeds the expected number of *de novo* mutations in the general population.[29] The presence of a statistically significant number of mutations in a gene in individuals with overlapping phenotypic features is currently the golden standard to establish causality between mutations in a specific gene and a particular human disorder.

Unsolved exomes

Unfortunately, a clear-cut pathogenic *de novo* mutation cannot be identified in all individuals with a sporadic phenotype in whom we perform trio-based WES. This can be due to technical issues including low sequencing quality or coverage in any of the trio samples or mis-mapping of reads to erroneous regions of the genome.[30] A mutation may be sequenced correctly in a sample but may not be called by the variant calling software or may be filtered out during

analysis as a candidate pathogenic mutation for various reasons. For instance, disease-causing mutations can be mistakenly interpreted as benign due to misleading *in silico* predictions on evolutionary conservation or effect. Additionally, mutations may be filtered out if they are found in a gene known to be involved in a different phenotype. The inability to identify a genetic mutation that could explain a phenotype may also be due to the phenotype having a different form of inheritance than hypothesized, leading to erroneous prioritization of candidate mutations. Incomplete penetrance, variable expressivity of the phenotype or non-Mendelian forms of inheritance may also impede the correct prioritization of candidate mutations. Finally, we regularly detect mutations predicted to be damaging in genes that are biologically plausible for the underlying phenotype in the patient but that have not been previously linked to disease. It is unclear in these situations whether the mutation is truly causative or not for the phenotype of the patient and it is essential to find at least one additional patient with a similar genetic lesion and clinical phenotype to replicate the finding and establish causality. This is particularly troublesome for rare and ultra-rare phenotypes, but a novel way to find such individuals is emerging from genetic matchmaking platforms such as Matchmaker exchange[31] or GeneMatcher[32] (www.matchmakerexchange.org and www.genematcher.org, respectively), which enable easy data sharing.

As an example of how pathogenic mutations can be missed in WES data, I tried to identify the disease-causing gene for Neu-Laxova syndrome (NLS), a rare autosomal recessive disorder, by sequencing three individuals affected by this disorder and two obligate carriers of the disease allele. Despite meticulous analysis and sequencing multiple samples, I initially failed to identify any pathogenic mutations that could explain the phenotype. It was only after *PHGDH* was found to be involved in this disorder,[33] that I identified pathogenic mutations in two other genes, *PSAT1* and *PSPH*, within the same biological pathway.[34] Retrospectively, I observed that three of the disease-causing variants were present in the sequencing data but were not called by the variant calling software: one was found in direct proximity to a known SNP, one was a 5 nucleotide indel and one was a deletion of at least 4.5 kb. Additionally, NLS was genetically more heterogeneous that initially hypothesized and the families included in our study had mutations in three different genes *PHGDH, PSAT1* and *PSPH*, which complicated the search of a mutation in the same gene in two individuals. Finally, homozygous mutations in all three genes had been previously identified in children with neurological alterations, a clinical phenotype distinct from NLS that leads to prenatal or perinatal lethality due to severe malformations affecting the central nervous system.

Targeted enrichment methods, such as WES, may fail to target the region where a disease-causing mutation is located such as non-coding areas of the genome. Recent studies have shown that *de novo* mutations in non-coding regions of the genome can cause human disease.[26] However, even if detected by WGS, these mutations often remain difficult to pinpoint as pathogenic amongst

the dozens of benign *de novo* variants that arise throughout the genome in each generation. As WGS becomes routine, variation in the non-coding regions of the genome of thousands of healthy individuals will be catalogued databases such as Kaviar[35] (db.systemsbiology.net/kaviar) and the Genome Aggregation Database (gnomad.broadinstitute.org), which will undoubtedly help in detecting pathogenic non-coding mutations.

Mosaicism and missing mutations

Failure to detect genetic mutations underlying a phenotype may be due to postzygotic occurrence of pathogenic mutations. Mosaicism for a pathogenic mutation may go undetected due to low levels or even absence of the mutation in the sampled tissue undergoing genetic analysis.

Some monogenic disorders, such as Cornelia de Lange syndrome (CdLS) have a relatively high prevalence of mosaicism for pathogenic mutations.[36] There are several possible explanations for this observation, which may also apply to other disorders. First, a large proportion of pathogenic *de novo* mutations in CdLS may arise postzygotically rather than in the germline, for yet unknown reasons. Second, individuals with CdLS may present a high incidence of revertant mosaicism, in which a cell carrying a germline pathogenic mutation mutates back to a wild-type allele.[37] Cells carrying a disease-causing mutation for CdLS may have a growth disadvantage compared to wild-type cells, which would result in the outgrowth of mutant cells reverting to a wild-type allele. This has been suggested to be a likely mechanism in CdLS, particularly for cells within the hematopoietic compartment.[36] Finally, another explanation may lie, not in a higher incidence of mosaicism, but in a higher probability for survival in individuals with CdLS caused by postzygotic mutations compared to individuals with CdLS caused by germline mutations.

Some disorders are caused exclusively by mosaic *de novo* mutations arising during early embryonic development, such as Proteus syndrome or Sturge-Weber syndrome.[38,39] The disease-causing mutations underlying these phenotypes have never been observed as germline events, suggesting either that they are incompatible with gamete differentiation if they arise during gametogenesis or that they are lethal when present constitutively in the zygote.[40] Interestingly, the pathogenic mutations responsible for Proteus syndrome and Sturge-Weber syndrome have been observed as somatic events driving cancer, which supports the hypothesis that they may block cellular differentiation either in gametogenesis or in early embryonic development.

Genetic mosaicism is difficult to detect using traditional methods for genetic analysis. For instance, Sanger sequencing lacks sensitivity to detect mutations present in less than 20% of cells within a sample. Prior to the development of NGS, clinical suspicion for mosaicism often guided the

identification of such mutations by meticulously analyzing chromatogram traces to identify varying levels of mutations among different samples from the same individual. As a result, our current knowledge and understanding of mosaicism is likely underestimating its contribution to human biology and disease. Mosaicism has been detected in disorders in which phenotypic changes are grossly observable or which affect tissues that can be obtained easily, such as the skin or blood. There is no reason, however, why other disorders lacking these characteristics would not be caused by mosaicism for genetic mutations.[41] With further improvements in NGS leading to a higher detection rate for these mutations, it is likely that we will see an explosion in this field and become aware of the contribution of mosaicism to many human diseases.

NGS to unveil mosaic mutations

Mosaicism can be detected by identifying deviations in the variant allele fraction (VAF) suggestive of a genetically heterogeneous population of cells within a sample. The digital nature of NGS provides sequence information for individual DNA molecules, which expands the range of detection for mosaicism that can be achieved with traditional methods. As shown in **Chapter 4**, high level mosaicism with mutations present in more than 60% of cells is difficult to distinguish from true heterozygous germline mutations when sequenced by Sanger sequencing. Furthermore, NGS is indispensable to detect low level mosaicism with mutations present in less than 20% of cells, as Sanger sequencing most often fails to detect these mutations.[42]

For NGS, the accuracy in the detection of mosaicism is dependent on the sequencing quality and sequencing coverage. Detection of high level mosaicism (present in more than 20% of cells) is based on the identification of mutations with a VAF deviating from the 50:50 ratio observed for heterozygous mutations. Higher sequencing coverage leads to a more precise representation of the allelic ratio in the biological sample, which is essential to detect whether a mutation presents a statistically significant deviation from the 50:50 ratio. Similarly, high sequencing coverage increases the sensitivity for detection of low-level mosaicism, which is based on the identification of mutations with low VAF that are statistically significantly higher than the background sequencing error. WGS and WES are usually performed at less than 100-fold coverage which translates into a probability of less than 40% to detect a mutation present in 10% of cells. Targeted approaches can provide sequencing coverage of over 1000-fold which lowers the detection threshold considerably. The additional use of molecular tagging or unique molecular identifiers (UMIs) can significantly increase the sensitivity of these methods. For instance, in **Chapter 5,** the UMI used in smMIPs allows one to trace multiple sequencing reads back to an individual DNA molecule captured and detect mutations present in 0.1% of cells. UMIs can be used to guide the generation of a consensus sequence from that individual

molecule, thereby circumventing both PCR amplification errors during library preparation, as well as, to a certain degree, sequencing errors.[43] Detection of low level mosaicism can also be achieved with technologies that do not rely on increased sequencing coverage; for instance, digital droplet PCR can detect mutations with a VAF of 0.1%. Nevertheless, sequencing-based approaches for the detection of low-level mosaicism encounter the problem that below a certain threshold, very low level mosaicism can easily be mistaken for sequencing errors and *vice versa*. The detection of mosaicism may be furthered by developments in NGS, such as lowering the background sequencing error rate and improving algorithms for variant calling of mutations with low VAFs, in order to discriminate low-level mosaicism from background sequencing error. The implementation of long sequencing reads will enable phasing of mutations and detection of a third haplotype in the presence of somatic mutations. Finally, single-cell sequencing represents a promising technique for the study of extremely low-level mosaicism as it is based on sequencing individual cells within a mixed population.[44]

Mosaicism can also be detected by comparing sequencing data from different tissues. Multiple samples per individual are occasionally collected in the clinical setting, particularly if mosaicism is suspected, but this is often avoided to prevent patient discomfort. Thus, implementing minimally invasive sampling of DNA from multiple tissues as a regular practice in clinical genetics would facilitate the routine detection of mosaicism. For instance, extracting cell-free DNA from plasma can represent a non-invasive form of sampling for which the original tissue serving as the source of DNA can be identified by methylation studies[45] or by examining nucleosome positioning[46].

As the detection of mosaicism improves and its role in disease becomes more widely understood, an effort needs to take place to improve the interpretation of the pathogenicity of mosaicism for disease-causing mutations. While standards and guidelines for the interpretation of genetic variants present in the germline have been released by the American College of Medical Genetics and the Association for Molecular Pathology,[47] non-germline variants pose additional challenges in their interpretation. The timing of a mutation, the tissues in which it is present and its frequency within those tissues can have an effect on the resulting phenotype. As such, these features should be investigated and included in databases such as Human Gene Mutation Database and Clinvar, which serve as catalogues for pathogenic human genetic variation.

Postzygotic and somatic mutations in human biology and disease

As a result of the previously mentioned improvements in NGS, attention has been drawn to the unexpectedly high prevalence of postzygotic and somatic mutations in health as well as in disease. Mutations arise continuously in all tissues, from the first few cell divisions and all throughout life.

Stage		Description	Time
Early embryogenesis		Mutation arising in any cell of an organism between fertilization and primordial germ cell specification (Carnegie stages 1 to 6)	Day 1 to day 17
Somatic			
Prenatal mutation	Late embryogenesis	Mutation arising in any somatic cell of an organism after primordial germ cell specification until formation of all major organs (Carnegie stages 7 to 23)	Day 18 to day 56
	Fetal development	Mutation arising in any somatic cell of an organism during fetal development until birth	Week 9 to birth
Postnatal mutation	Early postnatal	Mutation arising in any somatic cell of an organism during infancy	From 0 to 3 years
	Late postnatal	Mutation arising in any somatic cell of an organism during childhood	From 3 years until puberty
	Adult	Mutation arising in any somatic cell of an organism after puberty	After puberty
Gonadal			
Early gametogenic		Mutation arising in any germ cell or germ cell precursor between specification of primordial germ cells and up until puberty	Day 56 of development to puberty
Late gametogenic		Mutation arising in any germ cell from puberty onwards	After puberty

Table 1. Terminology to describe temporal characteristics of mutations. [1] Note that the division between embryogenesis and fetal development is arbitrarily defined. During fetal development, organogenesis is concluded but the organs continue to grow and differentiate.

The terminology used to describe mutations in terms of location is standardized and divides mosaicism into somatic mosaicism affecting only somatic tissues, gonadal mosaicism affecting only gametes and gonosomal mosaicism in which both gonadal and somatic tissues are affected. This is in contrast with germline or constitutive mutations that are already present at the single cell stage and are therefore found in all cells of the organism. Constitutive mutations are present in the zygote before the first cell division, having occurred either before or after fertilization. As the terminology used to describe mutations in terms of time is less well specified, I propose the terms in Table 1 to be used to expand and complement the description of mutations to include timing characteristics.

In **Chapter 4**, we studied mutations arising in the context of gametogenesis and in early embryogenesis to determine the proportion of *de novo* mutations occurring postzygotically. By analyzing the VAF of mutations identified by trio-based WGS, I found that 6.5% of seemingly germline *de novo* mutations did not arise during gametogenesis in the parent but actually originate

from mutations during early embryogenesis in the offspring. This result is supported by recent observations based on comparative genetic studies of monozygotic twins by WGS,[48] genetic studies from multiple tissues from the same individual by WGS[49] and additional studies analyzing WGS data from parent-offspring trios.[50,51] Depending on when and where they arise, mutations can be transmitted to the offspring; mutations occurring prior to the specification of primordial germ cells will be present both in somatic cells and in the gametes, which entails that they can be passed on to the next generation.[49] In this same study, we came to the very conservative estimate that 0.1% of seemingly germline *de novo* mutations in the offspring in fact occurred during early embryogenesis in one of the parents, leading to the presence of the mutations both in germ cells as well as in blood. This finding was obtained by analyzing sequencing data from WGS to detect sequencing reads denoting low level mosaicism in one of the parents for all *de novo* mutations identified in the offspring. This analysis was based on a biased data set of *de novo* mutations, as only high quality mutations were included. By definition, this excludes mutations for which reads are identified in the parents which is likely to filter out most low-level mosaic mutations in parents. Because of this, we expected that the number 0.1% would be an underestimation of the true frequency of low-level parental mosaicism as the origin of *de novo* mutations. A recent study has examined this phenomenon in more detail and placed this number at 3.8%.[51] The combination of these results suggests that over 10% of *de novo* mutations do not arise during gametogenesis but during early embryonic development in the parents or in the offspring.

Although a substantial proportion of mutations arise early in life,[52] somatic mutations accumulate over time and mosaicism increases with aging. Different tissues present differences in the mutation rates as well as some degree of variation in the mutational profile, as a result of exposure to different mutational processes.[53–55] In **Chapter 5**, I study somatic mutations arising in blood throughout adulthood based on ultra-deep sequencing of DNA extracted from blood of 2,014 healthy individuals between 20 and 69 years of age. I detected an increase in the burden of somatic mutations with age, even though mutations in blood could be detected in individuals as young as 21 years old. The contribution of somatic mutations to physiology has not been studied in detail, but mosaicism leads to variability among cells within a tissue, which may well contribute to health and tissue adaptability.[40] Nevertheless, mutations occasionally have an effect on cells, leading to changes in cell behavior and ultimately to disease[40].

Role of timing of *de novo* mutations in human disease

Recurrence risk

As shown in **Chapter 4** and mentioned previously, although most *de novo* mutations arise during gametogenesis and are transmitted to the offspring

as germline events, more than 10% of *de novo* mutations occur as mosaic events. Approximately 4% of *de novo* mutations occur during early embryogenesis in the parent and are transmitted as a constitutive event to the offspring[51], while close to 7% occur during early embryogenesis in the offspring. Mutations arising during early embryogenesis can lead to gonosomal mosaicism entailing that a significant proportion of gametes may carry the mutation, which translates into a higher risk for transmitting this mutation to multiple offspring. This is exemplified by previous reports of families with clinically unaffected parents having two or more children affected by a disorder caused by a *de novo* mutation[56–59].

Genetic counseling received by parents of an individual with a disease caused by a pathogenic *de novo* mutation includes an estimation of the risk of recurrence of the phenotype in subsequent offspring close to 1%.[60] However, gonosomal mosaicism in the parent, evidenced by the presence of the mutation in blood and transmission of the mutation to the offspring, entails a higher risk of recurrence.[60] In contrast, the risk of recurrence of a mutation that occurred during early embryogenesis in the offspring is equal to the population risk for this mutation, which is often many orders of magnitude lower than 1%. Put together, this implies that over 10% of couples with a child affected by a disorder caused by a *de novo* mutation are receiving inaccurate genetic counseling on their recurrence risk, which has significant implications for subsequent reproductive decisions.

The high prevalence of early embryonic events as the origin of *de novo* mutations supports routine assessment for the presence of mosaicism both in the offspring as well as in the parents as part of the diagnosis of disorders caused by pathogenic *de novo* mutations. The identification of mosaicism either in blood one of the parents or in the affected offspring or the detection of the pathogenic mutation in paternal sperm would allow a more accurate estimation and a stratification of the risk of recurrence, which would provide personalized genetic counseling and informed reproductive decisions[60].

Mutation timing shapes the resulting phenotype

The timing of a *de novo* mutation not only influences the risk of recurrence for disease, but also has a direct effect on the percentage of affected cells within tissues and within the organism. In the context of pathogenic mutations, the proportion of mutated cells within a tissue or organism shapes the resulting disease phenotype.[61] The important role of timing in determining the effect of genetic mutations is supported by the existence of healthy individuals with mosaicism for pathogenic *de novo* mutations. This situation poses an interesting question: when does the occurrence of a pathogenic mutation no longer cause disease?[62] Parents of individuals with developmental disorders, presenting parental gonosomal mosaicism for the pathogenic mutation are for the most part clinically unaffected. However, there are reports of mosaic parents

presenting clinical features related to their offspring's disorder due to mosaicism for the pathogenic mutation in relevant tissues. For instance, a study recently reported on the father of an individual with Costello syndrome, in whom gonosomal mosaicism for his son's disease-causing mutation was identified.[63] The *HRAS* mutation, resulting in a G12S substitution, had a VAF close to 50% in the proband and of 8% in a buccal swab of the father. While the proband presented with a severe clinical picture characteristic of Costello syndrome, his mosaic father had relatively mild clinical features with patchy skin and hair abnormalities and mild developmental delay. Collectively, these findings suggest, first, that a pathogenic *de novo* mutation may or may not cause disease in an individual, depending on its timing. Second, that a pathogenic *de novo* mutation may cause different phenotypes depending on the timing at which it occurs. There are several genes in which mutations can lead to different disorders depending on the time in which they arise (see Figure 1 for an example of multiple phenotypes caused by HRAS G12S mutations). Notably, over 90 of these, which I refer to as onco-developmental genes (see Supplementary Tables S1 and S2), have been found to cause developmental disorders when mutated in the germline and have been implicated in cancer when mutated somatically.

Temporal pleiotropy of a gene may stem from it being expressed and having a role at different moments of human life. If one considers a temporal window as an interval of time during which a gene is expressed and plays a role in a cell, tissue or organism, a gene's temporal windows are the moments in which gene dysfunction can lead to a phenotype at the cell, tissue or organism level. Genes with a transitory role during development have a temporal window of pathogenicity after which their dysfunction no longer disrupts development. In healthy individuals with gonosomal mosaicism for pathogenic mutations in such genes, either the mutation occurred late enough so as to not affect relevant tissues or it arose after the closure of the window for gene function and, therefore, for gene pathogenicity. However, if a gene has several temporal windows encompassing distinct time periods throughout life, the dysfunction of a gene can still be disruptive at different intervals and thereby lead to different phenotypes (see Figure 2). For instance, germline mutations in genes of the RAS/MAPK pathway leading to Noonan syndrome are present in those individuals throughout life. These genes are functionally active during embryonic development and therefore, genetic dysfunction during this temporal window is accompanied by congenital malformations. However, mutations in these genes during postnatal life can result in development of leukemia, suggesting a role for these genes in hematopoiesis within a posterior temporal window. Therefore, one might expect mutations in genes of the RAS/MAPK pathway in individuals with Noonan syndrome to also have consequences in this postnatal window. Indeed, individuals with Noonan syndrome have subclinical hematological alterations, including mild activation of RAS/MAPK signaling and decreased apoptosis of

circulating hematopoietic progenitors,[64] as well as spontaneous growth of colonies from peripheral blood cells.[65] Furthermore, Noonan syndrome associates with a mild and transitory myeloproliferative disorder and leukemia during childhood.[66] Mutations in onco-developmental genes can lead both to developmental abnormalities and oncogenesis, indicating that these genes have several temporal windows for pathogenicity. This, in turn, suggests that onco-developmental genes are likely to have several temporal windows for gene expression and function. While a mutation in an onco-developmental gene may cause a clinical phenotype in the first temporal window it disrupts, the very same mutation may disrupt gene function in a posterior temporal window and lead to disease within a different temporal window for pathogenicity (Figure 2A and 2B).

SETBP1 as a model gene to study the timing of *de novo* mutations

During my PhD, I have used germline and somatic mutations in a gene called *SETBP1* as a model to study the temporal pleiotropy of

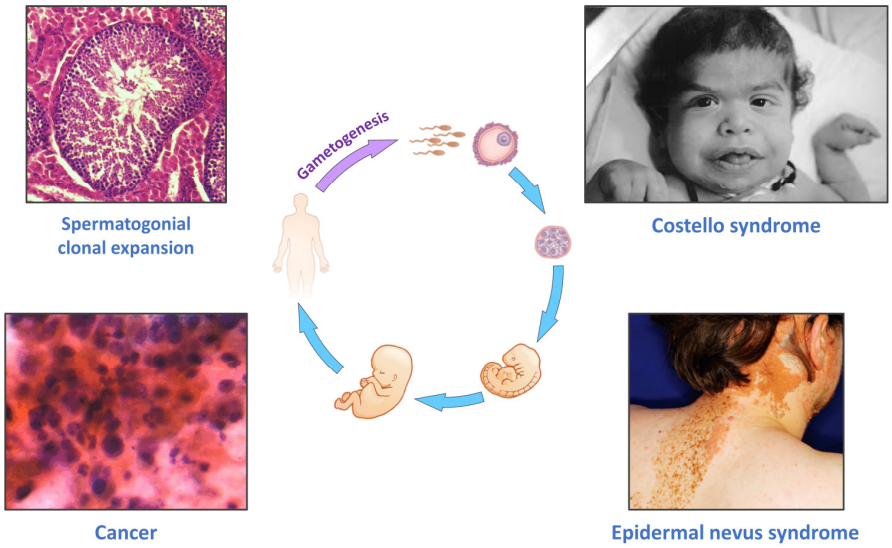


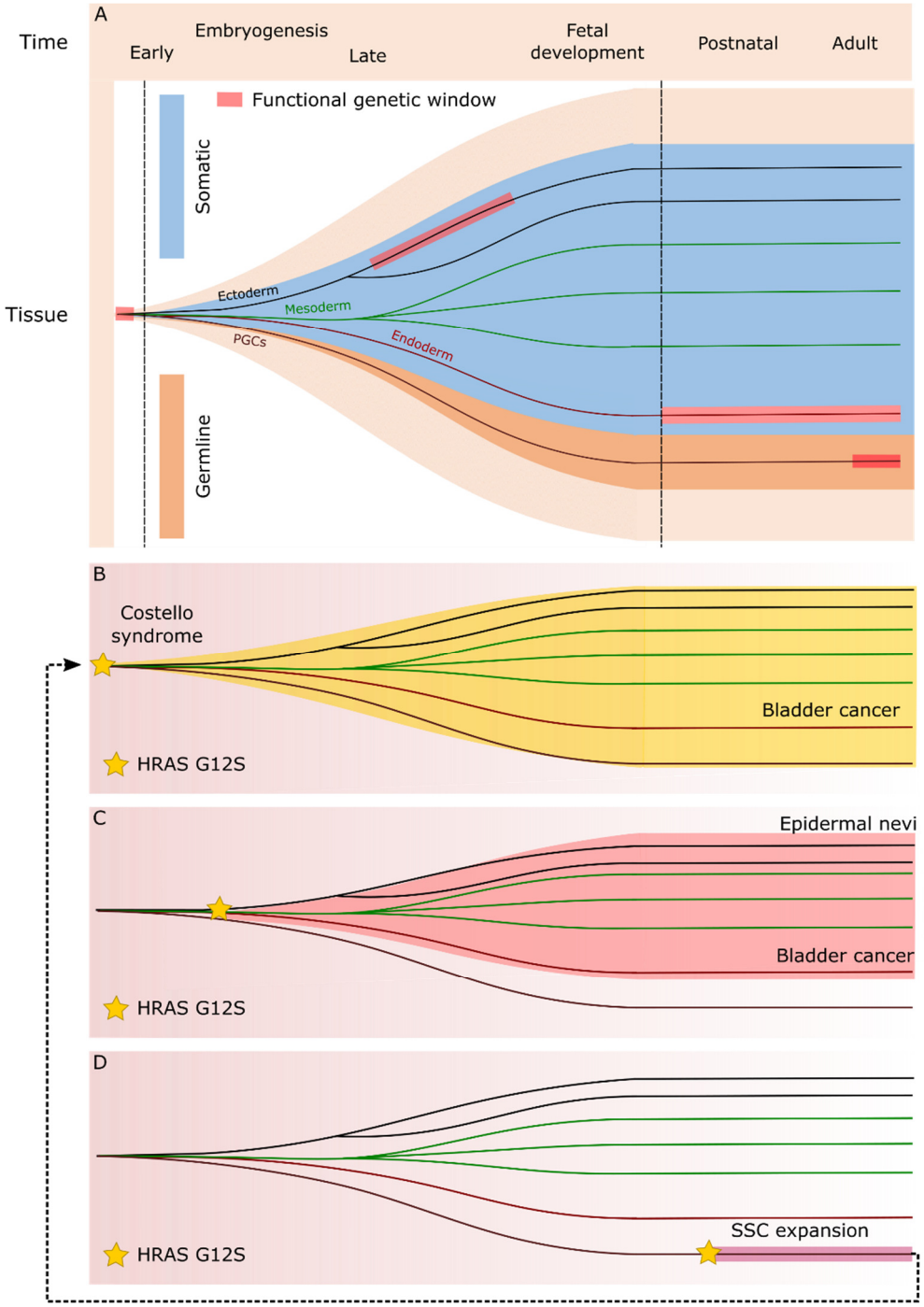
Figure 1. Pleiotropy of HRAS G12S mutations. The timing of *HRAS* mutations leading to HRAS G12S influences the resulting phenotype. Mutations present in the germline lead to Costello syndrome, a developmental disorder characterized by intellectual disability, characteristic facial features, heart abnormalities, skin laxity and joint flexibility. Mutations arising in embryogenesis have been identified in epidermal nevus and woolly hair syndrome. Remarkably, HRAS G12S mutations arising in embryogenesis have also been involved in recurrent urothelial cancer. Somatic HRAS G12S mutations have been observed in head and neck cancer, as well as urothelial and skin cancer. Finally, somatic mutations HRAS G12S arising in spermatogonial stem cells during spermatogenesis have been shown to lead to clonal expansion of spermatogonial stem cells. As a result of this clonal expansion of mutant spermatogonial stem cells, Costello syndrome is a paternal age effect disorder with elevated prevalence in the (aged) human population.

onco-developmental genes. Germline *de novo* mutations in a 12 bp coding region of *SETBP1* cause Schinzel-Giedion Syndrome (SGS), a rare and severe developmental disorder.[3] SGS is characterized by typical facial features, multiple congenital malformations, severe neurodevelopmental alterations and an increased incidence of malignancies, particularly neuroectodermal tumors. Remarkably, identical mutations arising during adulthood in hematopoietic stem cells have been implicated in myeloid malignancies.[67]

Parental origin and timing of de novo SETBP1 mutations in SGS

A recent report included the case of a woman who gave birth to a child with SGS and was found to have gonosomal mosaicism for the *SETBP1* mutation causative for the disease in her child.[68] The mother did not present with SGS but had a clinical history of neonatal malignant sacrococcygeal teratoma, one of the most frequent tumors in this disease. The mutation likely occurred in the mother at a time when she no longer presented the developmental phenotype resulting in congenital malformations characteristic for SGS, but could still develop a teratoma. This observation indicates that mosaicism for a mutation in an onco-developmental gene may cause disease other than the expected developmental phenotype.

In unpublished work, I examined the parental origin and timing of *de novo* mutations in *SETBP1* found in individuals with SGS to trace back the origin of these disease-causing mutations and potentially identify early embryonic *SETBP1* mutations. We used smMIPS to perform trio-based ultra-deep sequencing of the *SETBP1* hotspot in 21 trios with SGS to detect mosaicism either in the affected proband or in the parents. All *SETBP1* mutations were present as constitutive events in the proband and we could not identify the pathogenic mutation in any of the parental blood samples. Due to the small size of our cohort, the absence of postzygotic events among the 21 *SETBP1* mutations studied is not statistically significant and it is therefore difficult to interpret whether this result has biological meaning. However, one possibility could be that pathogenic *SETBP1* mutations arising in early embryogenesis do not cause SGS but a different phenotype, such as isolated teratoma, and are therefore absent from our cohort. Using a long-read sequencing platform, we then sequenced a region of 10 kb surrounding the *SETBP1* hotspot in 22 individuals with SGS in order to phase the pathogenic *de novo* mutation and genotype the mutant allele. We subsequently performed trio-based genotyping of this region to determine whether the *de novo* mutation occurred in the maternal or paternal allele. We were able to identify the parental allele in which the mutation occurred for 17 individuals, of which 8 originated on the maternal allele and 9 on the paternal allele. This


A
Time
Early
Embryogenesis
Late
Fetal development
Postnatal
Adult
Functional genetic window
Somatic
Tissue
Ectoderm
Mesoderm
PGCs
Endoderm
Germline
B
Costello syndrome
Bladder cancer
HRAS G12S
C
Epidermal nevi
Bladder cancer
HRAS G12S
D
SSC expansion
HRAS G12S

◀ **Figure 2. Temporal windows for gene functions.** A functional genetic window is considered here as an interval within time and space in which a gene is expressed, translated into a protein and has a role in a specific cell, tissue or organism. Some genes may have several windows encompassing distinct tissues and time periods throughout life. In panel A, different functional genetic windows are shown in pink for *HRAS*, which is likely to play a role in different tissues and from early embryonic development to adulthood. In B, germline *HRAS* G12S mutations are accompanied by congenital malformations. However, the *HRAS* mutation is present in all cells of the organism throughout life in individuals with Costello syndrome (shown in yellow). Therefore, germline *HRAS* mutations may also have consequences in this posterior functional genetic window, such as high risk of developing cancer. It is not known whether spermatogonial stem cell (SSCs) expansion occurs in men with germline *HRAS* mutations, but the mutation should not confer increased fitness to SSCs, as all cells in the organism carry the mutation. In C, an *HRAS* mutation arising during early embryogenesis does not give rise to Costello syndrome but to mosaic skin manifestations, such as epidermal nevi, and has been associated with increased risk of bladder cancer. In D, an *HRAS* mutation arises in spermatogenesis, granting an advantage to mutant cells. Production of mutant sperm cells and transmission of the mutation to the offspring causes Costello syndrome (in B, shown by the dotted line).

suggests that 53% of mutations in our cohort occurred during paternal gametogenesis and 47% originated during maternal gametogenesis. Since approximately 80% of *de novo* mutations arising during gametogenesis are of paternal origin, this distribution with similar frequency of paternal versus maternal origin of mutations is unexpected. This finding could reflect increased mutagenesis or selection for *SETBP1* mutant oocytes during maternal gametogenesis, although the reason behind this observation may once more lie in the small size of our cohort.

Functional thresholds for germline and somatic mutations

In **Chapter 6**, I present a genetic and functional comparison of overlapping germline and somatic *de novo* mutations arising in *SETBP1*. The downstream consequences observed in cell lines derived from individuals with pathogenic germline *SETBP1* mutations were comparable to those reported previously in hematopoietic stem cells expressing mutant SETBP1.[69] These observations include similar biochemical and cellular alterations resulting from *SETBP1* mutations arising in different temporal windows such as increased levels of SET protein and enhanced cell proliferation. However, further downstream, the biochemical and cellular consequences of germline and somatic *SETBP1* mutations may diverge as these mutations disrupt the function of the gene within different temporal windows. A gene may be involved in different processes throughout different time points of human life. This may be intrinsic to the gene itself, due to moonlighting activity, in which the same protein product from a gene has two distinct functions, or because of changes in levels of gene expression or expressed isoforms. Furthermore, each temporal window is characterized by a distinct cellular, tissue and organism context. Thus, the expression of the same mutation at different time points may lead to different

phenotypes due to distinct roles of the protein and a biological context which may differ considerably between one time point and the other.

Additionally, in **Chapter 6**, I describe differences between germline and somatic *SETBP1* mutations observed despite their overlap. As a group, somatic mutations in *SETBP1* identified in cancer are on average more disruptive to SETBP1 function than those identified as germline events in SGS. This finding suggests that embryonic cells in a developing organism are more sensitive to *SETBP1* disruption than somatic cells. Obviating differences in mutational mechanisms in the germline compared to somatic cells, this finding exposes differences in pathogenicity between germline and somatic mutations. Organisms are likely more resilient to biological disruption arising somatically than to germline disruptions. For one, constitutive mutations are present in all cells of an organism while somatic mutations only affect a subset of cells within that organism. Furthermore, embryonic cells in a developing organism are more plastic than somatic cells in a fully developed organism, entailing that a weaker stimulus can already disturb the homeostasis of the former while a stronger stimulus is required to disturb the latter. Thus, mutations have to be functionally more disruptive to drive diseases such as cancer in somatic cells. This hypothesis is supported by the observation that in our SGS cohort, individuals with SGS who developed cancer had on average more disruptive *SETBP1* mutations than those who did not. The existence of different functional thresholds for germline and somatic mutations has been proposed for other genes[66]. For instance, somatic mutations in *PTPN11* observed in leukemia are rarely identified as germline events in Noonan syndrome and *vice versa*.[70] Somatic *PTPN11* mutations are functionally stronger than germline mutations and have been hypothesized to be lethal when present in the germline.[71] Although most *PTPN11* mutations observed in Noonan syndrome occur exclusively as germline events, individuals with germline *PTPN11* mutations which have been identified as somatic events in cancer have been suggested to be at an increased risk of developing cancer[70]. The observation that the type of *de novo* mutation influences the expression of a phenotype depending on the time implies the existence of timing-related differences in cell, tissue and organism resilience to genetic mutations underlying distinct germline and somatic functional thresholds.

Somatic consequences of germline SETBP1 mutations

Germline *SETBP1* mutations are associated with an increased risk of malignancy, mainly with a high incidence of neuroectodermal tumors. Remarkably, despite the fact that somatic mutations in *SETBP1* plays a role in the development of myeloid malignancies, leukemia does not occur frequently in individuals with SGS. Many possible explanations can be offered to explain this observation. As shown in **Chapter 5**, the presence of one cancer-driving mutation is not sufficient for the development of cancer. Mutations are acquired in a

stepwise fashion and it is known from studies in other tissues that stochastic loss of mutant clones can occur. Therefore, the development of cancer requires time for mutations to accumulate and become fixed in tissues. Individuals with SGS have reduced life expectancy, which results in less time to accumulate mutations, but also in a biological context that is inadequate for these mutations to drive the development of myeloid leukemia. Myeloid leukemia is a disorder typically diagnosed in individuals over the age of 60 and it is likely that age-related changes in hematopoietic stem cells and in the bone marrow niche play a role in its development. For instance, as shown in **Chapter 5**, mutations involved in leukemia are accumulated throughout life and some mutations, such as those observed in genes within the spliceosome, only expand in the context of an aged cellular background. Furthermore, the somatic acquisition of a cancer-driving mutation confers the mutated cell an advantage over other cells. In an individual with a germline mutation, all cells are mutated and therefore the presence of the mutation in a cell does not grant it an advantage over other cells in the organism. This would suggest that the presence of mosaicism for a pathogenic mutation as a result of a mutation during embryonic development would result in genetically distinct cell populations and allow such competition to take place in a mosaic individual. Additionally, the existence of functional thresholds for mutations implies that mutations involved in cancer are on average stronger than those observed in developmental disorders. Thus, at least some individuals with developmental disorders could have a pathogenic mutation able to disrupt embryonic development but not strong enough to drive the development of cancer.

Future perspectives

Identifying individuals at increased risk for cancer

Identical mutations in onco-developmental genes leading to developmental disorders as well as cancer highlight a genetic overlap between both conditions. Additionally, multiple studies have described a clinical correlation between developmental disorders and cancer showing that individuals with congenital abnormalities have a higher incidence of pediatric cancer compared to the general pediatric population.[72–75] Conversely, individuals with pediatric cancer also show high prevalence of morphological abnormalities compared to controls.[76] This clinical association hints at common pathological processes underlying developmental disorders and cancer, which can at least be partially explained by an overlap at the genetic level. One study found that close to 4% of individuals with pediatric cancer had a well-defined genetic disorder such as trisomy 21, a RASopathy or an overgrowth syndrome.[77] Considering the high prevalence of mosaicism arising during early embryogenesis, a significant proportion of pediatric cancer could be associated with mosaicism for a pathogenic mutation in an onco-developmental gene without an overt

developmental phenotype. For instance, mosaicism for pathogenic *NRAS* mutations has been identified in children with juvenile myelomonocytic leukemia without a developmental syndrome.[78] This suggests first, that onco-developmental genes may have different windows of pathogenicity during prenatal development and postnatal life; and second, that mutations arising after the closure of the window for pathogenicity during embryonic development and leading to subclinical mosaicism could contribute to the development of cancer.

Gonosomal mosaicism in parents of individuals with developmental disorders caused by mutations in a onco-developmental gene may place them at risk for developing cancer. This is exemplified by the clinical case of a woman with gonosomal mosaicism for a pathogenic *SETBP1* mutation who developed a teratoma.[68] Epidemiological studies have examined cancer risk in parents of children with congenital malformations, with mixed results. One study found no overall increased risk for cancer in parents of individuals with congenital malformations.[73] Another study confirmed this negative finding for parents of children with malformations in general, but identified an increased risk for parents of children born with oral clefts.[79] Gonosomal mosaicism is expected to be present in 4% of parents of individuals with a developmental disorder caused by a *de novo* mutation in a onco-developmental gene. Thus, individuals with gonosomal mosaicism for a mutation in onco-developmental gene could represent a yet-undetected subgroup of the population at increased risk for developing cancer.

The accumulation of mutations throughout life drives the development of cancer, which entails that mutations arising during embryogenesis could be a frequent mechanism contributing to cancer in the general population.[80] A recent study of pediatric cancer found that 0.3% of their cohort presented mosaic mutations in tumor suppressors *TP53* or *RB1*.[81] Furthermore, at least 7% of germline mutations in *TP53* in individuals with Li-Fraumeni syndrome occurred *de novo*,[82] suggesting that a fraction of these could be caused by postzygotic mutations. If this is the case, screening for cancer could be based on identifying this type of mutations by sequencing germline and cell-free DNA in plasma rather than searching for alterations in metabolic markers, histology or gross macroscopic changes.

Pharmacological treatment of developmental disorders

The genetic and biological overlap between developmental disorders and cancer could represent a window of opportunity for intervention and therapy. Several cancer drugs exist on the market that target specific molecular alterations resulting from mutations in onco-developmental genes. These drugs could therefore be repurposed for the treatment of developmental disorders. For instance, postnatal treatment of a mouse model of Kabuki syndrome with a histone deacetylase inhibitor, originally intended for the treatment of cancer,

rescued neurological deficits in this mouse.[83] Additionally, segmental overgrowth disorders caused by mutations in *PIK3CA* could be treated with drugs such as rapamycin, used in cancer treatment for the inhibition of downstream pathways.[84] Rapamycin has also been shown to improve vascular malformations caused by *TIE2* mutations both in mouse models and in humans.[85] These hopeful findings may put an end to the dogma that developmental disorders cannot be treated.

High throughput interpretation of disease-causing mutations

As NGS becomes integrated in routine medical testing, millions of genetic variants will be identified in the human genome and will need to be assessed for pathogenicity.[4] However, the interpretation of the functional consequences of genetic mutations has failed to follow a development comparable to that of NGS and most novel or rare mutations fall by default in the category of variants of unknown significance (VUS). For example, up to 10% of women undergoing genetic testing for *BRCA1/2* mutations receive a VUS result and then face the decision of whether or not to pursue irreversible surgical treatment without knowing the clinical significance of their variant.[86] Currently available *in silico* methods to predict the effect of mutations are often incorrect[87] and, as a result, novel approaches for high-throughput and accurate evaluation of mutations are essential to deal with the immense number of genetic variants identified by large scale NGS studies. Saturation Genome Editing (SGE) is a new high-throughput genome editing technique in which hundreds of mutations are generated *in vitro* and are studied in parallel in a single assay.[88] After mutagenesis, a mixed population of cells, each of them harboring one of the studied mutations, is subjected to a functional assay selected based on the gene of interest. A single experiment can be performed to assess the effect of mutations on mRNA levels and splicing, protein function, protein-protein interaction or more complex cellular phenotypes such as cell signaling or proliferation. Finally, sequencing on a NGS platform is used as a read out. SGE is a flexible and highly versatile method that allows for the analysis of the effect of hundreds of mutations in parallel. This has potential applications in the area of diagnostics, as SGE can be used to create a catalogue of mutations with a functional interpretation prior to their identification in a patient sample. Furthermore, SGE can be performed in different cell lines, which allows one to study the effects of different genetic mutations in a cell-specific context. During my PhD, I performed a pilot SGE experiment coupled to a cellular assay to observe the effect of 253 mutations in *HRAS* and 459 mutations in *SETBP1* on proliferation of K562 cells (see Figure 3). While gene editing by CRISPR-induced homology-directed repair was observed robustly, I was unable to see statistically significant differences in the effect of the introduced mutations on cell proliferation. The choice and setup of the functional assay used for cell selection after mutagenesis clearly represents a challenge and requires further optimization.

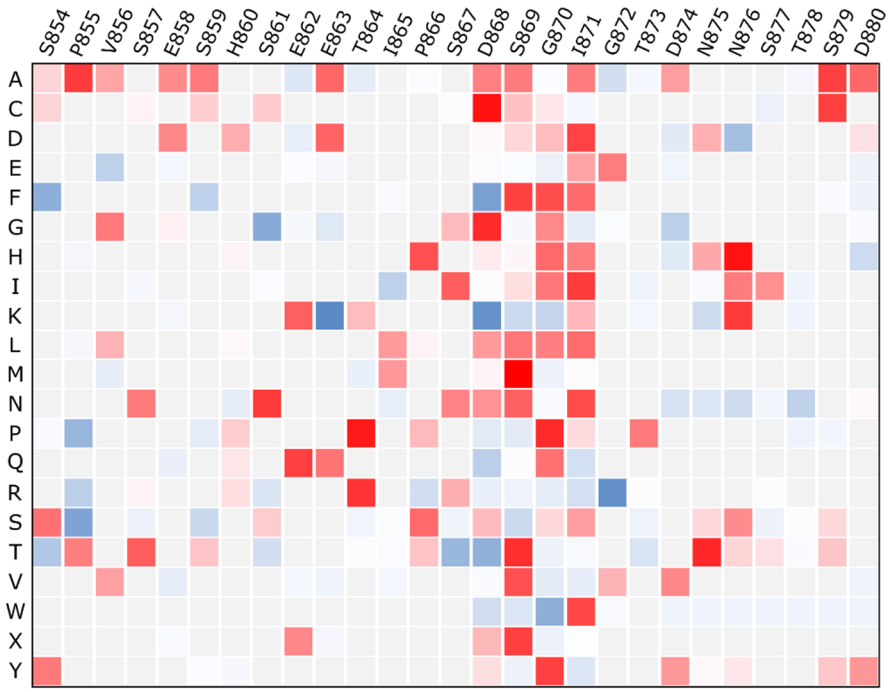


Figure 3. Effect of mutations introduced by Saturation Genome Editing in the *SETBP1* degron on proliferation of K562 cells, compared to wild-type cells. The mutated residue is indicated on the top of the figure, ranging from S854 to D880. Note that residues S854 to S867 and G872 to D880 underwent single nucleotide mutagenesis, while residues within the canonical degron of SETBP1 (D868 to I871) underwent trinucleotide mutagenesis to all possible codons. The amino acid to which the residue was mutated is indicated on the left side of the figure. Red indicates less cell counts while blue indicated more cell counts than cells in which the wild-type allele was introduced**.** Gray indicates no available data for the mutation at that position.

Conclusion

Over the last ten years, the field of human genetics has undergone a transformation fueled by the application of NGS in research and in the clinic. Trio-based sequencing, in particular, has enabled research on the biology and consequences of *de novo* mutations. Novel mutations arise between one generation and the next and continue to occur throughout life, giving rise to extraordinary genetic diversity among the population but also within a single individual. This proves the concept that "all the cells in our organism share the same genome" wrong. This genetic diversity is likely to play a role in human physiology and it contributes to sporadic human disease, both rare and common. While we are becoming aware of the role of *de novo* mutations in sporadic human disease, the significance of mosaicism in human disease is only starting to emerge. Mutations arise throughout prenatal development and postnatal life, illustrating that mosaicism is much more widespread than previously considered. As such, mosaicism may represent an important biological mechanism with implications on prenatal development, human physiology and evolution. Furthermore, mosaicism may explain certain aspects of human disease that continue to elude us, such as the link between developmental disease and cancer. As NGS technology continues to develop, we will gain further insight into the role of mosaicism in human biology and human disease, which will grant us a better understanding of the mechanism by which mutations arise and the effect of the expression of a mutation in the dynamic context of a cell, a tissue and a whole organism.

Web resources

Cancer Gene census: http://cancer.sanger.ac.uk/census (accessed on 16/12/2016)
Catalogue of Somatic Mutations in Cancer: http://cancer.sanger.ac.uk/cosmic
Developmental Disorder Genotype-Phenotype Database (DDG2P): https://decipher.sanger.ac.uk/about#downloads/data (accessed on 16/12/2016)
Human Gene Mutation Database: www.hgmd.cf.ac.uk

References

1. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* (2009). doi:10.1073/pnas.0910672106
2. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42,** 30–35 (2010).
3. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* **42,** 483–485 (2010).
4. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97,** 199–215 (2015).
5. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. a. Unlocking Mendelian disease using exome sequencing. *Genome Biol* **12,** 228 (2011).
6. Chiu, R. W. K. *et al.* Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A* **105,** 20458–20463 (2008).
7. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* **11,** 426–437 (2011).
8. Gagan, J. & Van Allen, E. M. Next-generation sequencing to guide cancer therapy. *Genome Med* **7,** 80 (2015).
9. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10,** 57–63 (2009).
10. Meaburn, E. & Schulz, R. Next generation sequencing in epigenetics: Insights and challenges. *Semin Cell Dev Biol* **23,** 192–199 (2012).
11. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol* **11,** R22 (2010).
12. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).
13. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
14. Michaelson, J. J. *et al.* Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151,** 1431–1442 (2012).
15. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47,** 822–826 (2015).
16. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* **48,** 935–939 (2016).
17. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
18. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12,** 628–640 (2011).
19. Sunyaev, S. R. Inferring causality and functional significance of human coding DNA variants. *Hum Mol Genet* **21,** R10–R17 (2012).
20. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).
21. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9,** e1003709 (2013).

22. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46,** 944–950 (2014).
23. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
24. Higurashi, N. *et al.* A human Dravet syndrome model from patient induced pluripotent stem cells. *Mol Brain* **6,** 19 (2013).
25. Kuechler, A. *et al.* Loss-of-function variants of SETD5 cause intellectual disability and the core phenotype of microdeletion 3p25.3 syndrome. *Eur J Hum Genet* **23,** 753–760 (2015).
26. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161,** 1012–1025 (2015).
27. McRae, J. F. *et al.* Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv* 1–39 (2016).
28. Stessman, H. A., Bernier, R. & Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* **156,** 872–877 (2014).
29. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* **19,** 1194–1196 (2016).
30. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12,** 745–755 (2011).
31. Philippakis, A. A. *et al.* The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum Mutat* **36,** 915–921 (2015).
32. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Hum Mutat* **36,** 928–930 (2015).
33. Shaheen, R. *et al.* Neu-laxova syndrome, an inborn error of serine metabolism, is caused by mutations in PHGDH. *Am J Hum Genet* **94,** 898–904 (2014).
34. Acuna-Hidalgo, R. *et al.* Neu-Laxova Syndrome Is a Heterogeneous Metabolic Disorder Caused by Defects in Enzymes of the L-Serine Biosynthesis Pathway. *Am J Hum Genet* **95,** 285–293 (2014).
35. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L. & Roach, J. C. Kaviar: An accessible system for testing SNV novelty. *Bioinformatics* **27,** 3216–3217 (2011).
36. Huisman, S. A., Redeker, E. J. W., Maas, S. M., Mannens, M. M. & Hennekam, R. C. M. High rate of mosaicism in individuals with Cornelia de Lange syndrome. *J Med Genet* **50,** 339–344 (2013).
37. Jongmans, M. C. J. *et al.* Revertant somatic mosaicism by mitotic recombination in dyskeratosis congenita. *Am J Hum Genet* **90,** 426–433 (2012).
38. Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* **365,** 611–9 (2011).
39. Shirley, M. D. *et al.* Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. *N Engl J Med* **368,** 1971–9 (2013).
40. Fernández, L. C., Torres, M. & Real, F. X. Somatic mosaicism: on the road to cancer. *Nat Rev Cancer* **16,** 43–55 (2015).
41. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat Rev Genet* **14,** 307–320 (2013).
42. Rohlin, A. *et al.* Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. *Hum Mutat* (2009). doi:10.1002/humu.20980
43. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* **23,** 843–854 (2013).
44. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14,** 681–91 (2013).
45. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci* **112,** E5503–E5512 (2015).
46. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164,** 57–68 (2016).
47. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and

the Association for Molecular Pathology. *Genet Med* **17,** 405–423 (2015).
48. Dal, G. M. *et al.* Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J Med Genet* **51,** 455–459 (2014).
49. Huang, A. Y. *et al.* Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res* **24,** 1311–1327 (2014).
50. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6,** 5969 (2015).
51. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48,** 126–133 (2015).
52. Finette, B. A. *et al.* Determination of hprt mutant frequencies in T-lymphocytes from a healthy pediatric population: statistical comparison between newborn, children and adult mutant frequencies, cloning efficiency and age. *Mutat Res Mol Mech Mutagen* **308,** 223–231 (1994).
53. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47,** 1402–1407 (2015).
54. Dolle, M. E. T., Snyder, W. K., Gossen, J. A., Lohman, P. H. M. & Vijg, J. Distinct spectra of somatic mutations accumulated with age in mouse heart and small intestine. *Proc Natl Acad Sci* **97,** 8403–8408 (2000).
55. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538,** 260–264 (2016).
56. Elalaoui, S. C. *et al.* Germinal mosaicism in Noonan syndrome: A family with two affected siblings of normal parents. *Am J Med Genet Part A* (2010). doi:10.1002/ajmg.a.33685
57. Natacci, F. *et al.* Germline mosaicism in achondroplasia detected in sperm DNA of the father of three affected sibs. *Am J Med Genet Part A* (2008). doi:10.1002/ajmg.a.32228
58. Helderman-van Den Enden, A. T. J. M. *et al.* Recurrence risk due to germ line mosaicism: Duchenne and Becker muscular dystrophy. *Clin Genet* (2009). doi:10.1111/j.1399-0004.2009.01173.x
59. Willemsen, M. *et al.* Familial Kleefstra syndrome due to maternal somatic mosaicism for interstitial 9q34.3 microdeletions. *Clin Genet* **80,** 31–38 (2011).
60. Campbell, I. M. *et al.* Parent of origin, mosaicism, and recurrence risk: Probabilistic modeling explains the broken symmetry of transmission genetics. *Am J Hum Genet* **95,** 345–359 (2014).
61. Papavassiliou, P. *et al.* The phenotype of persons having mosaicism for trisomy 21/Down syndrome reflects the percentage of trisomic cells present in different tissues. *Am J Med Genet Part A* **149A,** 573–583 (2009).
62. Morita, H. & Komuro, I. Somatic Mutations in Cerebral Cortical Malformations. *N Engl J Med* **371,** 2037–2038 (2014).
63. Sol-Church, K. *et al.* Male-to-male transmission of Costello syndrome: G12S HRAS germline mutation inherited from a father with somatic mosaicism. *Am J Med Genet Part A* **149A,** 315–321 (2009).
64. Timeus, F. *et al.* Functional evaluation of circulating hematopoietic progenitors in Noonan syndrome. *Oncol Rep* (2013). doi:10.3892/or.2013.2535
65. Gaipa, G. *et al.* Peripheral blood cells from children with RASopathies show enhanced spontaneous colonies growth in vitro and hyperactive RAS signaling. *Blood Cancer J* **5,** e324 (2015).
66. Tartaglia, M. *et al.* Diversity and functional consequences of germline and somatic PTPN11 mutations in human disease. *Am J Hum Genet* **78,** 279–90 (2006).
67. Piazza, R. *et al.* Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat Genet* **45,** 18–24 (2012).
68. Vendrell Bayona, T. *et al.* Somatic mosaic mutation in SETBP1 identified in the mother of a girl with Schinzel-Giedion syndrome. in *European Society for Human Genetics* P11.140 (2016).
69. Makishima, H. *et al.* Somatic SETBP1 mutations in myeloid malignancies. *Nat Genet* **45,** 942–6 (2013).
70. Jongmans, M. C. J. *et al.* Cancer risk in patients with Noonan syndrome carrying a PTPN11 mutation. *Eur J Hum Genet* (2011). doi:10.1038/ejhg.2011.37
71. Kratz, C. P. *et al.* The mutational spectrum of PTPN11 in juvenile myelomonocytic leukemia and Noonan syndrome/myeloproliferative disease. *Blood* **106,** 2183–5 (2005).

72. Agha, M. M. *et al.* Congenital abnormalities and childhood cancer. *Cancer* **103,** 1939–1948 (2005).
73. Bjorge, T., Cnattingius, S., Lie, R. T., Tretli, S. & Engeland, A. Cancer Risk in Children with Birth Defects and in Their Families: A Population Based Cohort Study of 5.2 Million Children from Norway and Sweden. *Cancer Epidemiol Biomarkers Prev* **17,** 500–506 (2008).
74. Narod, S. A., Hawkins, M. M., Robertson, C. M. & Stiller, C. A. Congenital anomalies and childhood cancer in Great Britain. *Am J Hum Genet* **60,** 474–85 (1997).
75. Durmaz, A. *et al.* The Association of minor congenital anomalies and childhood cancer. *Pediatr Blood Cancer* **56,** 1098–102 (2011).
76. Merks, J. H. M. *et al.* Prevalence and Patterns of Morphological Abnormalities in Patients With Childhood Cancer. *JAMA* **299,** 61–69 (2008).
77. Merks, J. H. M., Caron, H. N. & Hennekam, R. C. M. High incidence of malformation syndromes in a series of 1,073 children with cancer. *Am J Med Genet* (2005). doi:10.1002/ajmg.a.30603
78. Doisaki, S. *et al.* Somatic mosaicism for oncogenic NRAS mutations in juvenile myelomonocytic leukemia. *Blood* **120,** 1485–1488 (2012).
79. Zhu, J. L. *et al.* Do parents of children with congenital malformations have a higher cancer risk? A nationwide study in Denmark. *Br J Cancer* (2002). doi:10.1038/sj.bjc.6600488
80. Hafner, C., Toll, A. & Real, F. X. HRAS Mutation Mosaicism Causing Urothelial Cancer and Epidermal Nevus. *N Engl J Med* **365,** 1940–1942 (2011).
81. Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med* **373,** 2336–2346 (2015).
82. Gonzalez, K. D. *et al.* High frequency of de novo mutations in Li-Fraumeni syndrome. *J Med Genet* **46,** 689–693 (2009).
83. Bjornsson, H. T. *et al.* Histone deacetylase inhibition rescues structural and functional brain deficits in a mouse model of Kabuki syndrome. *Sci Transl Med* **6,** 256ra135-256ra135 (2014).
84. Lindhurst, M. J. M. J. *et al.* Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. *Nat Genet* **44,** 928–33 (2012).
85. Boscolo, E. *et al.* Rapamycin improves TIE2-mutated venous malformation in murine model and human subjects. *J Clin Invest* **125,** 3491–3504 (2015).
86. Murray, M. L., Cerrato, F., Bennett, R. L. & Jarvik, G. P. Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions. *Genet Med* **13,** 998–1005 (2011).
87. Masica, D. L. & Karchin, R. Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. *PLoS Comput Biol* **12,** 1–16 (2016).
88. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513,** 120–123 (2014)

Images from Figure 1

Image of testes from goo.gl/TcG3Ar (Creative commons license)
Image of individual with Costello syndrome from:
Johnson J.P. Costello syndrome: phenotype, natural history, differential diagnosis, and possible cause. *J Pediatr* **133**:441-8 (1998)
Reproduction with permission of Elsevier.
Image of individual with epidermal nevus syndrome from:
Groessler L. *et al.* , Phacomatosis pigmentokeratotica is caused by a postzygotic HRAS mutation in a multipotent progenitor cell *J Invest Dermatol* **133**:1998-2003 (2013)
Reproduction with permission of Elsevier.
Image of cancer cells from: goo.gl/JTRCz8 (Creative commons license)

Supplementary data

Gene	Developmental phenotype
AFF4	Cornelia De Lange-Like Syndrome
AKT1	Proteus Syndrome
AMER1	Osteopathia Striata With Cranial Sclerosis
AR	Spinal And Bulbar Muscular Atrophy
AR	Androgen Insensitivity Syndrome
ARID1A	Coffin-Siris Syndrome
ARID1B	Coffin Siris Syndrome
ARID1B	Intellectual disability, Autosomal Dominant 12
ASXL1	Bohring-Opitz Syndrome
ATRX	Alpha-Thalassemia Mental Retardation Syndrome X-Linked Non-Deletion Type
ATRX	Mental Retardation Syndromic X-Linked With Hypotonic Facies Syndrome
BCL11A	Intellectual Disability
BCOR	Microphthalmia Syndromic Type 2
BRAF	Cardiofaciocutaneous Syndrome
BRAF	Leopard Syndrome Type 3
BRAF	Noonan Syndrome Type 7
CBL	Noonan Syndrome-Like Disorder With Or Without Juvenile Meylomonocytic Leukemia
CCND2	Megalencephaly-Polymicrogyria-Polydactyly-Hydrocephalus Syndrome
CHD4	CHD4 Syndrome
CHD4	Syndromic Intellectual Disability With Or Without Congenital Heart Disease
CNOT3	CNOT3 Syndrome
COL1A1	COL1A1/2-Related Osteogenesis Imperfecta
COL1A1	Ehlers-Danlos Syndrome Type VIIA
COL1A1	Ehlers-Danlos Syndrome, Classic Type, COL1A1-Related
COL1A1	Osteogenesis Imperfecta Type I
COL1A1	Osteogenesis Imperfecta Type IIA
COL1A1	Osteogenesis Imperfecta Type III
COL1A1	Caffey Disease
COL2A1	Spondyloepimetaphyseal Dysplasia Strudwick Type
COL2A1	Achondrogenesis Type 2

Gene	Developmental phenotype
COL2A1	Kniest Dysplasia
COL2A1	Platyspondylic Lethal Skeletal Dysplasia Torrance Type
COL2A1	Rhegmatogenous Retinal Detachment Autosomal Dominant
COL2A1	Spondyloperipheral Dysplasia
COL2A1	Stickler Syndrome Type 1 Non-Syndromic Ocular
COL2A1	Spondyloepiphyseal Dysplasia Congenita
CREBBP	Rubinstein-Taybi Syndrome Type 1
CTCF	Intellectual Disability
CTNNB1	Mental Retardation, Autosomal Dominant 19
DDX3X	Intellectual Disability
DNMT3A	Overgrowth Syndrome With Intellectual Disability
ELN	ELN-Related Cutis Laxa
ELN	Supravalvar Aortic Stenosis
EP300	Rubinstein-Taybi Syndrome Type 2
EZH2	Weaver Syndrome 2
FGFR1	Encephalocraniocutaneous Lipomatosis
FGFR1	Osteoglophonic Dysplasia
FGFR1	Pfeiffer Syndrome
FGFR1	Idiopathic Hypogonadotropic Hypogonadism
FGFR1	Kallmann Syndrome Type 2
FGFR2	Acrocephalosyndactyly Type V
FGFR2	Apert Syndrome
FGFR2	Beare-Stevenson Cutis Gyrata Syndrome
FGFR2	Crouzon Syndrome
FGFR2	Jackson-Weiss Syndrome
FGFR2	Antley-Bixler Syndrome
FGFR2	Familial Scaphocephaly Syndrome
FGFR2	Lacrimo-Auriculo-Dento-Digital Syndrome
FGFR3	Achondroplasia
FGFR3	Crouzon Syndrome With Acanthosis Nigricans
FGFR3	Hypochondroplasia
FGFR3	Muenke Syndrome

Gene	Developmental phenotype
FGFR3	Thanatophoric Dysplasia Type 1
FGFR3	Thanatophoric Dysplasia Type 2
FGFR3	Camptodactyly Tall Stature And Hearing Loss Syndrome
FGFR3	Lacrimo-Auriculo-Dento-Digital Syndrome
FLT4	Milroy Disease
FOXP1	Mental Retardation With Language Impairment And Autistic Features
GATA2	Emberger Syndrome
GNAS	ACTH-Independent Macronodular Adrenal Hyperplasia
GNAS	*GNAS* Hyperfunction
GNAS	Albright Hereditary Osteodystrophy
GRIN2A	Epilepsy With Neurodevelopmental Defects
GRIN2A	Landau-Kleffner Syndrome
HOXA13	Hand-Foot-Genital Syndrome
HOXD13	Brachydactyly Type D
HOXD13	Brachydactyly Type E
HOXD13	Brachydactyly-Syndactyly Syndrome
HOXD13	Syndactyly Type 5
HOXD13	Synpolydactyly 1
HOXD13	VACTERL Association
HRAS	Congenital Myopathy With Excess Of Muscle Spindles
HRAS	Costello Syndrome
KAT6A	Mental Retardation, Autosomal Dominant 32
KAT6B	Genitopatellar Syndrome
KAT6B	Blepharophimosis/Intellectual Disability Phenotype Which Is Noonan-Like
KDM5C	Mental Retardation Syndromic X-Linked JARID1C-Related
KDM6A	Kabuki Syndrome 2
KIT	Human Piebaldism
KMT2A	Wiedemann-Steiner Syndrome
KMT2D	Kabuki Syndrome
KRAS	Cardiofaciocutaneous Syndrome
KRAS	Noonan Syndrome Type 3
LMNA	Hutchinson-Gilford Progeria Syndrome

Gene	Developmental phenotype
LMNA	Lethal Tight Skin Contracture Syndrome
LMNA	Emery-Dreifuss Muscular Dystrophy Type 2
LMNA	Cardiomyopathy Dilated Type 1A
LMNA	Cardiomyopathy Dilated With Hypergonadotropic Hypogonadism
LMNA	Familial Partial Lipodystrophy Type 2
LMNA	Heart-Hand Syndrome Slovenian Type
LMNA	Limb-Girdle Muscular Dystrophy Type 1B
LMNA	Mandibuloacral Dysplasia With Type A Lipodystrophy
LMNA	Muscular Dystrophy Congenital LMNA-Related
MAF	Cataract, Deafness, ID, Seizures, And Down Syndrome-Like Facies
MAF	Cataract Congenital Cerulean Type 4
MAF	Cataract Pulverulent Juvenile-Onset Maf-Related
MAFB	Multicentric Carpotarsal Osteolysis Syndrome
MAP2K1	Cardiofaciocutaneous Syndrome
MAP2K2	Cardiofaciocutaneous Syndrome
MAP3K1	46XY Sex Reversal 6
MED12	Lujan-Fryns Syndrome
MED12	Opitz-Kaveggia Syndrome
MITF	Waardenburg Syndrome Type 2 With Ocular Albinism
MITF	Waardenburg Syndrome Type 2A
MITF	Tietz Syndrome
MNX1	Currarino Syndrome
MYCN	Feingold Syndrome Type 1
MYH9	Macrothrombocytopenia With Progressive Sensorineural Deafness
MYH9	May-Hegglin Anomaly
MYH9	Sebastian Syndrome
MYH9	Deafness Autosomal Dominant Type 17
MYH9	Epstein Syndrome
MYH9	Fechtner Syndrome
NF1	Familial Spinal Neurofibromatosis
NF1	Neurofibromatosis Type 1
NF1	Neurofibromatosis-Noonan Syndrome

Gene	Developmental phenotype
NF1	Watson Syndrome
NKX2-1	Benign Hereditary Chorea
NKX2-1	Choreoathetosis, Hypothyroidism, And Neonatal Respiratory Distress
NOTCH2	Hajdu-Cheney Syndrome
NRAS	Noonan Syndrome Type 6
NSD1	Sotos Syndrome
NSD1	Weaver Syndrome
NSD1	Beckwith-Wiedemann Syndrome
PAX3	Craniofacial-Deafness-Hand Syndrome
PAX3	Waardenburg Syndrome, Type 1
PAX8	Congenital Hypothyroidism Non-Goitrous Type 2
PDGFRB	Familial Infantile Myofibromatosis
PDGFRB	Premature Aging Syndrome, Penttinen Type
PHF6	Boerjeson-Forssman-Lehmann Syndrome
PHOX2B	Central Hypoventilation Syndrome, Congenital, With or Without Hirschsprung Disease
PHOX2B	Neuroblastoma With Hirschsprung Disease
PIK3CA	CLOVES; Congenital Lipomatous Overgrowth, Vascular Malformations, And Epidermal Nevi
PIK3CA	Hemimegalencephaly PIK3CA
PIK3CA	Megalencephaly-Capillary Malformation-Polymicrogyria Syndrome, Somatic 3
PIK3R1	Short Syndrome
PPM1D	PPM1D Syndrome
PPP2R1A	Intellectual Disability
PRKAR1A	Acrodysostosis
PTCH1	Basal Cell Nevus Syndrome
PTCH1	Holoprosencephaly-7
PTEN	Bannayan-Zonana Syndrome
PTEN	Cowden Disease
PTEN	Lhermitte-Duclos Disease
PTEN	Proteus Syndrome
PTEN	Macrocephaly/Autism Syndrome
PTEN	VACTERL Association With Hydrocephalus

Gene	Developmental phenotype
PTPN11	Noonan Syndrome 1
PTPN11	Leopard Syndrome Type 1
RAD21	Cohesinopathy
RAF1	Noonan Syndrome 5
RET	Multiple Endocrine Neoplasia IIB
SET	SET syndrome
SETBP1	Schinzel-Giedion syndrome
SETBP1	Developmental and Expressive Language Delay
SMAD3	*SMAD3*-Related Loeys-Dietz Syndrome
SMAD4	Myhre Syndrome
SMAD4	Juvenile Polyposis Syndrome
SMAD4	Juvenile Polyposis/Hereditary Hemorrhagic Telangiectasia Syndrome
SMARCA4	Coffin Siris syndrome
SMARCA4	Rhabdoid Tumor Predisposition Syndrome 2
SMARCB1	Rhabdoid Predisposition Syndrome 1
SMO	Curry-Jones Syndrome
SOX2	AEG Syndrome
SOX2	Microphthalmia Syndromic Type 3
TBX3	Ulnar-Mammary Syndrome
TCF12	Coronal Craniosynostosis
TGFBR2	Loeys-Dietz Syndrome
TGFBR2	*TGFBR2*-Related Loeys-Dietz Syndrome
TP63	Ectrodactyly-Ectodermal Dysplasia-Cleft Lip/Palate Syndrome Type 3
TP63	Acro-Dermato-Ungual-Lacrimal-Tooth Syndrome
TP63	Ankyloblepharon-Ectodermal Defects-Cleft Lip/Palate
TP63	Ectodermal Dysplasia Rapp-Hodgkin Type
TP63	Limb-Mammary Syndrome
TP63	Non-Syndromic Orofacial Cleft Type 8
TP63	Split-Hand/Foot Malformation Type 4
TSC1	Tuberous Sclerosis Type 1
TSC2	Lymphangioleiomyomatosis
TSC2	Tuberous Sclerosis Type 2

Gene	Developmental phenotype
TSHR	Hyperthyroidism, Familial Gestational
WT1	Denys-Drash Syndrome
WT1	Frasier Syndrome Frasier Syndrome Frasier Syndrome

Supplementary Table S1. Developmental disorders resulting from germline or postzygotic *de novo* mutations in onco-developmental genes. This list was obtained by overlapping the list of genes involved in developmental phenotypes in the Developmental Disorder Genotype-Phenotype Database (DDG2P) from the Deciphering Developmental Disorders study and a list of genes mutated somatically in cancer, obtained from the Cancer Gene census from the Catalogue of Somatic Mutations in Cancer database.

▶ **Supplementary Table S2.** Genes with overlapping germline and somatic mutations in onco-developmental genes observed in developmental disorders and in cancer, respectively. Germline and postzygotic pathogenic mutations leading to developmental disorders were extracted from the Human Gene Mutation Database. Recurrent somatic missense and truncating mutations in cancer were obtained from the Catalogue of Somatic Mutations in Cancer database by filtering for protein residues with 10 or more missense mutations and genes with 10 or more stop mutations throughout the coding sequence of the gene. Both sets of mutations were overlapped to identify missense and loss-of-function mutations leading to developmental disorders when arising in the germline or postzygotically and involved in cancer when arising somatically. XLR: X-linked recessive; XLD: X-linked dominant; Act: activating; LoF: loss-of-function; DN: dominant negative; Mis: missense; ?: uncertain; T: translocation; M: missense; N: nonsense; D: deletion; F: frameshift, S: Splice site; A: amplification; O: Other.

Gene	Developmental phenotype	Type of inheritance	Germline mut.	Somatic mut.
AFF4	Cornelia De Lange-Like Syndrome	Dominant	Act	T
AKT1	Proteus Syndrome	Mosaic	Act	M
AMER1	Osteopathia Striata With Cranial Sclerosis	Dominant	LoF	F, D, N, M
AR	Spinal And Bulbar Muscular Atrophy	XLR	DN	M
AR	Androgen Insensitivity Syndrome	XLR	LoF	M
ARID1A	Coffin-Siris Syndrome	Dominant	LoF	M, N, F, S, D, T
ARID1B	Coffin Siris Syndrome	Dominant	LoF	M, F, N, O
ARID1B	Intellectual disability, Autosomal Dominant 12	Dominant	LoF	M, F, N, O
ASXL1	Bohring-Opitz Syndrome	Dominant	LoF	F, N, M
ATRX	Alpha-Thalassemia Mental Retardation Syndrome	XLR	LoF	M, F, N
ATRX	Mental Retardation Syndromic X-Linked With Hypotonic Facies	XLR	LoF	M, F, N
BCL11A	Intellectual Disability	Dominant	Mis	T
BCOR	Microphthalmia Syndromic Type 2	XLD	LoF	F, N, S, T
BRAF	Cardiofaciocutaneous Syndrome	Dominant	Act	M, T, O
BRAF	Leopard Syndrome Type 3	Dominant	Act	M, T, O
BRAF	Noonan Syndrome Type 7	Dominant	Act	M, T, O
CBL	Noonan Syndrome-Like Disorder With Or Without Juvenile Meylomonocytic Leukemia	Dominant	Act	T, M, S, O
CCND2	Megalencephaly-Polymicrogyria-Polydactyly-Hydrocephalus Syndrome	Dominant	Act	T
CHD4	CHD4 Syndrome	Dominant	LoF	M, F, N
CHD4	Syndromic Intellectual Disability With Or Without Congenital Heart Disease	Dominant	LoF	M, F, N
CNOT3	CNOT3 Syndrome	Dominant	LoF	M, N, F
COL1A1	COL1A1/2-Related Osteogenesis Imperfecta	Dominant	DN	T
COL1A1	Ehlers-Danlos Syndrome Type VIIA	Dominant	DN	T
COL1A1	Ehlers-Danlos Syndrome, Classic Type, COL1A1-Related	Dominant	DN	T
COL1A1	Osteogenesis Imperfecta Type I	Dominant	DN	T
COL1A1	Osteogenesis Imperfecta Type IIA	Dominant	DN	T
COL1A1	Osteogenesis Imperfecta Type III	Dominant	DN	T
COL1A1	Caffey Disease	Dominant	?	T
COL2A1	Spondyloepimetaphyseal Dysplasia Strudwick Type	Dominant	Mis	F, M, N, T
COL2A1	Achondrogenesis Type 2	Dominant	DN	F, M, N, T

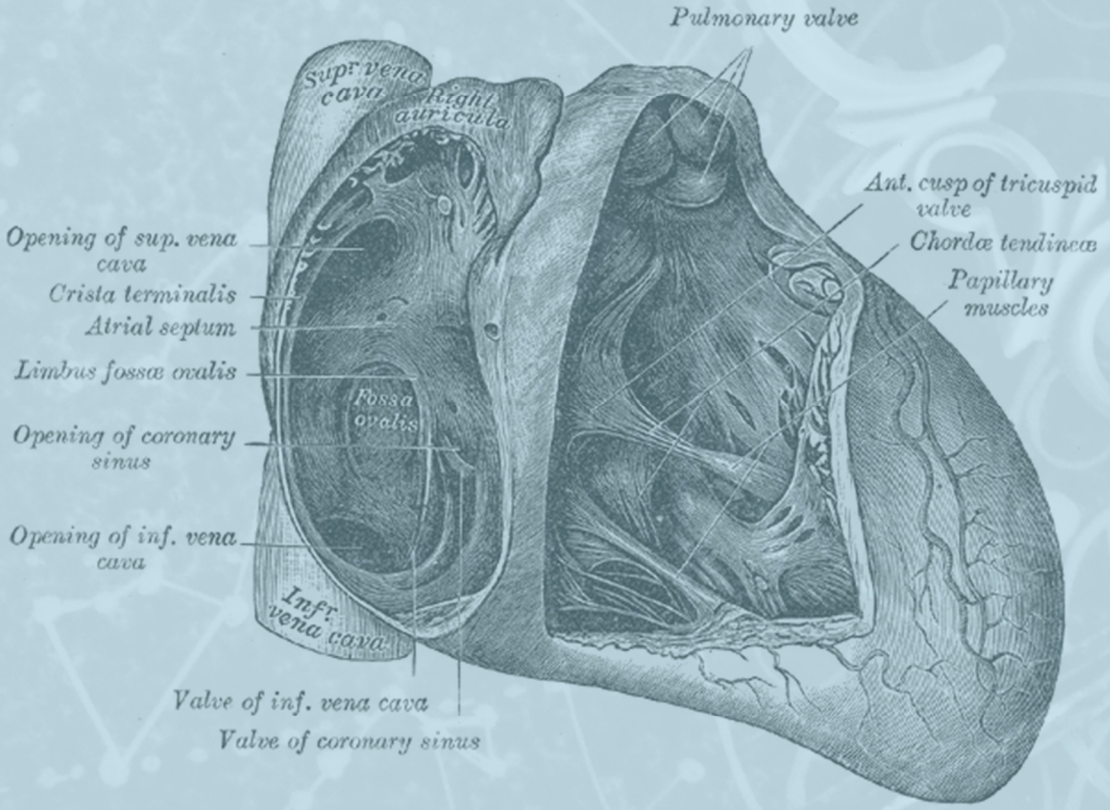
Gene	Developmental phenotype	Type of inheritance	Germline mut.	Somatic mut.
COL2A1	Kniest Dysplasia	Dominant	DN	F, M, N, T
COL2A1	Platyspondylic Lethal Skeletal Dysplasia Torrance Type	Dominant	LoF	F, M, N, T
COL2A1	Rhegmatogenous Retinal Detachment Autosomal Dominant	Dominant	LoF	F, M, N, T
COL2A1	Spondyloperipheral Dysplasia	Dominant	LoF	F, M, N, T
COL2A1	Stickler Syndrome Type 1 Non-Syndromic Ocular	Dominant	LoF	F, M, N, T
COL2A1	Spondyloepiphyseal Dysplasia Congenita	Dominant	?	F, M, N, T
CREBBP	Rubinstein-Taybi Syndrome Type 1	Dominant	LoF	T, N, F, M, O
CTCF	Intellectual Disability	Dominant	LoF	M, N
CTNNB1	Mental Retardation, Autosomal Dominant 19	Dominant	LoF	M, O, T
DDX3X	Intellectual Disability	XLR	Mis	M, N, F
DDX3X	Intellectual Disability	XLD	LoF	M, N, F
DNMT3A	Overgrowth Syndrome With Intellectual Disability	Dominant	LoF	M, F, N, S
ELN	ELN-Related Cutis Laxa	Dominant	LoF	T
ELN	Supravalvar Aortic Stenosis	Dominant	LoF	T
EP300	Rubinstein-Taybi Syndrome Type 2	Dominant	LoF	T, N, F, M, O
EZH2	Weaver Syndrome 2	Dominant	Mis	M
FGFR1	Encephalocraniocutaneous Lipomatosis	Mosaic	Act	T
FGFR1	Osteoglophonic Dysplasia	Dominant	Act	T
FGFR1	Pfeiffer Syndrome	Dominant	Act	T
FGFR1	Idiopathic Hypogonadotropic Hypogonadism	Dominant	LoF	T
FGFR1	Kallmann Syndrome Type 2	Dominant	LoF	T
FGFR2	Acrocephalosyndactyly Type V	Dominant	Act	M
FGFR2	Apert Syndrome	Dominant	Act	M
FGFR2	Beare-Stevenson Cutis Gyrata Syndrome	Dominant	Act	M
FGFR2	Crouzon Syndrome	Dominant	Act	M
FGFR2	Jackson-Weiss Syndrome	Dominant	Act	M
FGFR2	Antley-Bixler Syndrome	Dominant	?	M
FGFR2	Familial Scaphocephaly Syndrome	Dominant	?	M
FGFR2	Lacrimo-Auriculo-Dento-Digital Syndrome	Dominant	?	M
FGFR3	Achondroplasia	Dominant	Act	M, T
FGFR3	Crouzon Syndrome With Acanthosis Nigricans	Dominant	Act	M, T

Gene	Developmental phenotype	Type of inheritance	Germline mut.	Somatic mut.
FGFR3	Hypochondroplasia	Dominant	Act	M, T
FGFR3	Muenke Syndrome	Dominant	Act	M, T
FGFR3	Thanatophoric Dysplasia Type 1	Dominant	Act	M, T
FGFR3	Thanatophoric Dysplasia Type 2	Dominant	Act	M, T
FGFR3	Camptodactyly Tall Stature And Hearing Loss Syndrome	Dominant	?	M, T
FGFR3	Lacrimo-Auriculo-Dento-Digital Syndrome	Dominant	?	M, T
FLT4	Milroy Disease	Dominant	Mis	0
FOXP1	Mental Retardation With Language Impairment And Autistic Features	Dominant	LoF	T
GATA2	Emberger Syndrome	Dominant	LoF	M
GNAS	ACTH-Independent Macronodular Adrenal Hyperplasia	Mosaic	Act	M
GNAS	Gnas Hyperfunction	Dominant	Act	M
GNAS	Albright Hereditary Osteodystrophy	Dominant	LoF	M
GRIN2A	Epilepsy With Neurodevelopmental Defects	Dominant	LoF	M, N, F, O
GRIN2A	Landau-Kleffner Syndrome	Dominant	LoF	M, N, F, O
HOXA13	Hand-Foot-Genital Syndrome	Dominant	LoF	T
HOXD13	Brachydactyly Type D	Dominant	?	T
HOXD13	Brachydactyly Type E	Dominant	?	T
HOXD13	Brachydactyly-Syndactyly Syndrome	Dominant	?	T
HOXD13	Syndactyly Type 5	Dominant	?	T
HOXD13	Synpolydactyly 1	Dominant	?	T
HOXD13	VACTERL Association	Dominant	?	T
HRAS	Congenital Myopathy With Excess Of Muscle Spindles	Dominant	Act	M
HRAS	Costello Syndrome	Dominant	Act	M
KAT6A	Mental Retardation, Autosomal Dominant 32	Dominant	LoF	T
KAT6B	Genitopatellar Syndrome	Dominant	DN	T
KAT6B	Blepharophimosis/Intellectual Disability Phenotype (Noonan-Like)	Dominant	LoF	T
KDM5C	Mental Retardation Syndromic X-Linked JARID1C-Related	XLR	LoF	N, F, S, M
KDM6A	Kabuki Syndrome 2	XLD	LoF	D, N, F, S
KIT	Human Piebaldism	Dominant	LoF	M, O
KMT2A	Wiedemann-Steiner Syndrome	Dominant	LoF	T, O
KMT2D	Kabuki Syndrome	Dominant	LoF	N, F, M
KRAS	Cardiofaciocutaneous Syndrome	Dominant	Act	M

Gene	Developmental phenotype	Type of inheritance	Germline mut.	Somatic mut.
KRAS	Noonan Syndrome Type 3	Dominant	Act	M
LMNA	Hutchinson-Gilford Progeria Syndrome	Dominant	Act	T
LMNA	Lethal Tight Skin Contracture Syndrome	Dominant	Act	T
LMNA	Emery-Dreifuss Muscular Dystrophy Type 2	Dominant	LoF	T
LMNA	Cardiomyopathy Dilated Type 1A	Dominant	?	T
LMNA	Cardiomyopathy Dilated With Hypergonadotropic Hypogonadism	Dominant	?	T
LMNA	Familial Partial Lipodystrophy Type 2	Dominant	?	T
LMNA	Heart-Hand Syndrome Slovenian Type	Dominant	?	T
LMNA	Limb-Girdle Muscular Dystrophy Type 1B	Dominant	?	T
LMNA	Mandibuloacral Dysplasia With Type A Lipodystrophy	Dominant	?	T
LMNA	Muscular Dystrophy Congenital LMNA-Related	Dominant	?	T
MAF	Cataract, Deafness, Intellectual Disability, Seizures, And A Down Syndrome-Like Facies	Dominant	Mis	T
MAF	Cataract Congenital Cerulean Type 4	Dominant	?	T
MAF	Cataract Pulverulent Juvenile-Onset MAF-Related	Dominant	?	T
MAFB	Multicentric Carpotarsal Osteolysis Syndrome	Dominant	Act	T
MAP2K1	Cardiofaciocutaneous Syndrome	Dominant	Act	M
MAP2K2	Cardiofaciocutaneous Syndrome	Dominant	Act	M
MAP3K1	46XY Sex Reversal 6	Dominant	LoF	N, F, M, O, S
MED12	Lujan-Fryns Syndrome	XLR	?	M, S, O
MED12	Opitz-Kaveggia Syndrome	XLR	?	M, S, O
MITF	Waardenburg Syndrome Type 2 With Ocular Albinism	Dominant	LoF	A
MITF	Waardenburg Syndrome Type 2A	Dominant	LoF	A
MITF	Tietz Syndrome	Dominant	?	A
MNX1	Currarino Syndrome	Dominant	LoF	T
MYCN	Feingold Syndrome Type 1	Dominant	LoF	A
MYH9	Macrothrombocytopenia With Progressive Sensorineural Deafness	Dominant	LoF	T
MYH9	May-Hegglin Anomaly	Dominant	LoF	T
MYH9	Sebastian Syndrome	Dominant	LoF	T
MYH9	Deafness Autosomal Dominant Type 17	Dominant	?	T
MYH9	Epstein Syndrome	Dominant	?	T
MYH9	Fechtner Syndrome	Dominant	?	T
NF1	Familial Spinal Neurofibromatosis	Dominant	LoF	D, M, N, F, S, O
NF1	Neurofibromatosis Type 1	Dominant	LoF	D, M, N, F, S, O

Gene	Developmental phenotype	Type of inheritance	Germline mut.	Somatic mut.
NF1	Neurofibromatosis-Noonan Syndrome	Dominant	LoF	D, M, N, F, S, O
NF1	Watson Syndrome	Dominant	LoF	D, M, N, F, S, O
NKX2-1	Benign Hereditary Chorea	Dominant	LoF	A
NKX2-1	Choreoathetosis, Hypothyroidism, And Neonatal Respiratory Distress	Dominant	LoF	A
NOTCH2	Hajdu-Cheney Syndrome	Dominant	Act	N, F, M
NRAS	Noonan Syndrome Type 6	Dominant	Act	M
NSD1	Sotos Syndrome	Dominant	LoF	T
NSD1	Weaver Syndrome	Dominant	LoF	T
NSD1	Beckwith-Wiedemann Syndrome	Dominant	?	T
PAX3	Craniofacial-Deafness-Hand Syndrome	Dominant	LoF	T
PAX3	Waardenburg Syndrome, Type 1	Dominant	LoF	T
PAX8	Congenital Hypothyroidism Non-Goitrous Type 2	Dominant	LoF	T
PDGFRB	Familial Infantile Myofibromatosis	Dominant	Act	T
PDGFRB	Premature Aging Syndrome, Penttinen Type	Dominant	Act	T
PHF6	Boerjeson-Forssman-Lehmann Syndrome	XLR	LoF	F, N, S, M
PHOX2B	Central Hypoventilation Syndrome, Congenital, With Or Without Hirschsprung Disease	Dominant	?	M, F
PHOX2B	Neuroblastoma With Hirschsprung Disease	Dominant	?	M, F
PIK3CA	Cloves: Congenital Lipomatous Overgrowth, Vascular Malformations, And Epidermal Nevi	Mosaic	Act	M
PIK3CA	Hemimegalencephaly PIK3CA	Mosaic	Act	M
PIK3CA	Megalencephaly-Capillary Malformation-Polymicrogyria Syndrome, Somatic 3	Mosaic	Act	M
PIK3R1	Short Syndrome	Dominant	Mis	M, F, O
PPM1D	PPM1D Syndrome	Dominant	LoF	A, M, N, F
PPP2R1A	Intellectual Disability	Dominant	DN	M
PRKAR1A	Acrodysostosis	Dominant	Act	T, M, N, F, S
PTCH1	Basal Cell Nevus Syndrome	Dominant	LoF	M, N, F, S
PTCH1	Holoprosencephaly-7	Dominant	?	M, N, F, S
PTEN	Bannayan-Zonana Syndrome	Dominant	LoF	D, M, N, F, S
PTEN	Cowden Disease	Dominant	LoF	D, M, N, F, S
PTEN	Proteus Syndrome	Mosaic	LoF	D, M, N, F, S
PTEN	Macrocephaly/Autism Syndrome	Dominant	?	D, M, N, F, S
PTEN	VACTERL Association With Hydrocephalus	Dominant	?	D, M, N, F, S

Gene	Developmental phenotype	Type of inheritance	Germline mut.	Somatic mut.
PTPN11	Noonan Syndrome 1	Dominant	Act	M
PTPN11	Leopard Syndrome Type 1	Dominant	Mis	M
RAD21	Cohesinopathy	Dominant	LoF	M, N, F
RAF1	Noonan Syndrome 5	Dominant	Act	T
RET	Multiple Endocrine Neoplasia IIB	Dominant	Act	T, M, N, F
SET	Set Syndrome	Dominant	LoF	T
SETBP1	Schinzel-Giedion Midface Retraction Syndrome	Dominant	Act	M, T
SETBP1	Developmental And Expressive Language Delay	Dominant	LoF	M, T
SMAD3	*SMAD3*-Related Loeys-Dietz Syndrome	Dominant	LoF	M
SMAD4	Myhre Syndrome	Dominant	Act	D, M, N, F
SMAD4	Juvenile Polyposis Syndrome	Dominant	LoF	D, M, N, F
SMAD4	Juvenile Polyposis/Hereditary Hemorrhagic Telangiectasia Syndrome	Dominant	LoF	D, M, N, F
SMARCA4	Coffin Siris	Dominant	LoF	F, N, M, S
SMARCA4	Rhabdoid Tumor Predisposition Syndrome 2	Dominant	LoF	F, N, M, S
SMARCB1	Rhabdoid Predisposition Syndrome 1	Dominant	LoF	D, N, F, S
SMO	Curry-Jones Syndrome	Mosaic	Act	M
SOX2	AEG Syndrome	Dominant	LoF	A
SOX2	Microphthalmia Syndromic Type 3	Dominant	LoF	A
TBX3	Ulnar-Mammary Syndrome	Dominant	LoF	M, N, F, O
TCF12	Coronal Craniosynostosis	Dominant	LoF	T
TGFBR2	Loeys-Dietz Syndrome	Dominant	LoF	M, F, N
TGFBR2	TGFBR2-Related Loeys-Dietz Syndrome	Dominant	LoF	M, F, N
TP63	Ectrodactyly-Ectodermal Dysplasia-Cleft Lip/Palate Syndrome Type 3	Dominant	LoF	M, N, T
TP63	Acro-Dermato-Ungual-Lacrimal-Tooth Syndrome	Dominant	?	M, N, T
TP63	Ankyloblepharon-Ectodermal Defects-Cleft Lip/Palate	Dominant	?	M, N, T
TP63	Ectodermal Dysplasia Rapp-Hodgkin Type	Dominant	?	M, N, T
TP63	Limb-Mammary Syndrome	Dominant	?	M, N, T
TP63	Non-Syndromic Orofacial Cleft Type 8	Dominant	?	M, N, T
TP63	Split-Hand/Foot Malformation Type 4	Dominant	?	M, N, T
TSC1	Tuberous Sclerosis Type 1	Dominant	LoF	D, M, N, F, S
TSC2	Lymphangioleiomyomatosis	Dominant	LoF	D, M, N, F, S
TSC2	Tuberous Sclerosis Type 2	Dominant	LoF	D, M, N, F, S
TSHR	Hyperthyroidism, Familial Gestational	Dominant	Act	M
WT1	Denys-Drash Syndrome	Dominant	DN	D, M, N, F, S, T



Human heart.
Anatomy of the Human Body by Henry Gray & Henry Vandyke Carter (1918)

Chapter 8

Summary of this thesis

Nederlandse Samenvatting

Resumen en Español

Acknowledgements

List of publications

Curriculum Vitae

RIMLS portfolio

Summary of this thesis

All the information needed for the development, growth, life and reproduction of a human being is encoded in the human genome and is transmitted from parents to their offspring. Differences in DNA sequence among individuals are called genetic variants and most genetic variants in the genome of any individual were inherited from his or her parents. However, a fraction of these variants only occurred for the first time in that individual and are therefore considered to be new or *de novo* mutations. Like genetic mutations in general, *de novo* mutations can be benign, neutral or damaging and they play an important role in evolution, diversity and disease in humans. In this thesis, I study the time at which novel mutations arise throughout human life and examine the effect of timing on the consequences of novel mutations.

The mechanisms underlying the occurrence and genomic distribution of new mutations are reviewed in detail in **Chapter 2**. Novel mutations arise when DNA damage fails to be repaired, which can happen at any time and in any cell in the organism. As a result, novel mutations arise throughout human life, from mutations occurring in the sperm or egg cell during gametogenesis, giving rise to *de novo* mutations in the germline of the offspring, to mutations in somatic tissues during adult life.

Novel mutations are an important cause of human disease, particularly sporadic human phenotypes and severe disorders affecting fitness or fertility, such as severe pediatric disease. **Chapter 3** presents the identification of germline *de novo* mutations in *THRA* as the cause of a novel thyroid hormone resistance syndrome accompanied by developmental defects and alterations in growth. The results from this study illustrate the role that *de novo* mutations play in sporadic and severe pediatric disorders. Furthermore, the identification of the disease-causing mutations in this sporadic phenotype was enabled by exome sequencing and comparative genetic analysis of two unrelated individuals. This study highlights the importance of the development of NGS to our understanding of the role of *de novo* mutations in human disease, particularly sporadic developmental disorders.

De novo mutations most often arise during gametogenesis or prior to the first cell division of the zygote, but they can occur postzygotically and lead to mosaicism. Indeed, a number of human developmental disorders are caused by *de novo* mutations arising postzygotically and leading to genetic mosaicism. In **Chapter 4**, we use different NGS techniques to examine a set of seemingly germline *de novo* mutations and determine the contribution of postzygotic events to *de novo* mutations. In this study, we identified that 7% of *de novo* mutations which seemed to be present as constitutive events in the offspring

were in fact present as mosaic mutations suggesting that they had occurred postzygotically. Furthermore, several *de novo* mutations identified in individuals in our cohort were found back in blood in one of the parents as low level mosaic mutations, supporting that these mutations were not *de novo* in the offspring but had occurred during embryonic development in one of the parents. Together, these results indicate that the occurrence of *de novo* mutations is not limited to gametogenesis but that an important proportion of *de novo* mutations arise postzygotically. Genetic mosaicism resulting from mutations occurring during embryogenesis is common and, depending on their timing, postzygotic mutations can be transmitted to the next generation. This suggests that our genomes may be much more dynamic than previously considered.

The occurrence of new mutations in humans is not limited to the early stages of life and mutations continue to arise postnatally and during adult life. In **Chapter 5**, we investigate the timing of novel mutations in somatic tissues throughout life by examining somatic mutations during human hematopoiesis. In this study, we focus on mutations leading to positive selection of hematopoietic stem cells and driving clonal expansion. These mutations are known as clonal hematopoiesis driver mutations and are often present as low-level mosaic mutations in blood. In order to detect these somatic mutations, we developed an assay combining high-throughput and ultra-sensitive sequencing to screen the blood of individuals between 20 and 69 years of age. Clonal hematopoiesis driver mutations were detected in blood of 3% of individuals between 20 and 30 years of age. Furthermore, we observed an exponential increase in the frequency in clonal hematopoiesis driver mutations in blood with age, surpassing 20% of individuals between 60 and 69 years of age. The findings from this study support that the occurrence of somatic mutations in stem cells followed by the positive selection and clonal evolution of mutant stems cells is a universal mechanism occurring throughout human life.

The timing of a new mutation influences the consequences of this mutation. For instance, the same mutation can lead to completely different phenotypes, depending on the time at which it arises, as shown in **Chapter 6**. This chapter explores the functional consequences of overlapping germline and somatic mutations in *SETBP1*, leading to Schinzel-Giedion syndrome and to myeloid leukemia, respectively. In this study, we identified parallelisms between the downstream effects of overlapping germline and somatic mutations in *SETBP1*, including biochemical and cellular consequences. Furthermore, despite the overlap between germline and somatic *SETBP1* mutations in Schinzel-Giedion syndrome and myeloid leukemia, the mutation spectrum is different in both conditions. This stems from the fact that although *SETBP1* mutations with weak functional consequences are observed frequently as germline events in Schinzel-Giedion syndrome, they are rarely identified as somatic events in myeloid leukemia. This suggests that different functional thresholds exist for germline and somatic mutations with malignancy being driven strictly by strongly disruptive

mutations in *SETBP1*, while prenatal development is disrupted in Schinzel-Giedion syndrome by strong or mild mutations in this gene.

Finally, **Chapter 7** offers an in-depth discussion on the findings of this thesis, including the use of NGS to detect and study both *de novo* germline and somatic mutations and the role of the timing of novel mutations in shaping their consequences.

The work presented in this thesis shows that mutations arise constantly, between one generation and the next but also throughout life. This continuous occurrence of mutations leads to extraordinary genetic diversity between individuals but is also at the origin of the existence of genetically different populations of cells within a single human being. While the occurrence of novel mutations represents an important biological phenomenon in humans with a role on prenatal development, physiology and evolution, novel mutations also contribute to different forms of human disease ranging from rare and severe developmental disorders to adult-onset diseases such as cancer. In the last few years, NGS has played a central role in the identification and study of novel mutations, driving research in human genetics and quickly translating results into clinical practice. The work in this thesis supports that in addition to the detection of mutations, NGS can be used as a tool to identify the timing of mutations in order to include the dimension of time in the interpretation of mutations and their possible consequences. For instance, *de novo* mutations have different recurrence risks depending on the exact timing at which they occurred, which can be determined by meticulous analysis with NGS methods. Furthermore, the timing of a mutation can also shape the resulting phenotype, as the same mutation can be involved in different disorders depending on its timing. Future work focusing on the role of timing of somatic and germline mutations will provide us with a better understanding of the effect of the expression of a mutation in the dynamic context of a cell, a tissue and a whole organism.

Nederlandse samenvatting

Alle informatie die nodig is voor de ontwikkeling, groei, het leven en de voortplanting van een mens is vastgelegd in het menselijk genoom en wordt doorgegeven door ouders aan hun nageslacht. Verschillen in het DNA tussen individuen worden genetische varianten genoemd. Alhoewel de meeste genetische varianten in het genoom van een individu overerft zijn van zijn of haar ouders, ontstaat een klein aantal varianten voor de eerste keer in het individu, en deze worden daarom beschouwd als nieuwe varianten, ook wel *de novo* mutaties genoemd. Net als genetisch varianten kunnen *de novo* varianten goedaardig, neutraal of schadelijk zijn, en spelen ze een rol in evolutie, diversiteit en ziekte in de mens. In dit proefschrift bestudeer ik wanneer deze *de novo* mutaties ontstaan gedurende de menselijk levensloop en onderzoek ik het effect van het ontstaansmoment op de consequenties van deze nieuwe mutaties.

De mechanismen die ten grondslag liggen aan het ontstaan en de genomisch spreiding van nieuwe mutaties worden in detail beschouwd in **hoofdstuk 2**. Nieuwe mutaties kunnen ontstaan wanneer beschadigd DNA niet goed gerepareerd wordt, iets wat op ieder moment kan voorkomen, in iedere cel van het organisme. Het gevolg hiervan is dat nieuwe mutaties gedurende het gehele menselijke leven kunnen ontstaan, beginnend bij de ontwikkeling van de sperma of eicel, maar ook in somatische weefsels gedurende het volwassen leven.

Nieuw ontstane mutaties zijn een belangrijke oorzaak van menselijke ziektes, met name sporadische humane fenotypes en ernstige aandoeningen met een effect op "fitness" of vruchtbaarheid zoals ernstige aangeboren aandoeningen. **Hoofdstuk 3** toont dat kiembaan *de novo* mutaties in het gen *THRA* de oorzaak zijn van een nieuw schildklier-hormoon-resistentie syndroom dat gepaard gaat met ontwikkelingsdefecten en groeiveranderingen. De resultaten van deze studie illustreren de bijdrage van *de novo* mutaties aan sporadische en ernstige kindergeneeskundige aandoeningen. Bovendien werden deze bevindingen gedaan middels "exome sequencing" en het vergelijken van de genetische profielen van twee niet-gerelateerde individuen. Deze studie onderstreept het belang van NGS voor de ontwikkeling van ons begrip van de rol van *de novo* mutaties in menselijke ziektes in het algemeen, en voor sporadische ontwikkelingsstoornissen in het bijzonder.

De novo mutaties ontstaan meestal gedurende de gametogenese of voor de eerste celdeling van de zygoot, maar kunnen ook post-zygotisch ontstaan en leiden tot mozaïcisme. Inderdaad wordt een aantal humane ontwikkelingsziekten veroorzaakt door *de novo* mutaties die post-zygotisch ontstaan en resulteren in

genetisch mozaïcisme. In **hoofdstuk 4** gebruiken we verschillende NGS technieken om een aantal ogenschijnlijke *de novo* mutaties te onderzoeken, en te bepalen wat het aandeel is van post-zygotische gebeurtenissen hierin. In deze studie vonden we dat 7% van *de novo* mutaties die ogenschijnlijk aanwezig waren als constitutieve gebeurtenissen in het nageslacht, feitelijk aanwezig waren als mozaïeke mutaties, suggererend dat deze post-zygotisch waren ontstaan. Bovendien was het zo dat een aantal *de novo* mutaties die we vonden in individuen in ons cohort, ook werd teruggevonden in het bloed van de ouders als laaggradige mozaïeke mutaties. Dit bevestigde dat deze mutaties niet nieuw ontstaan waren in het nageslacht, maar waren ontstaan gedurende de embryonale ontwikkeling in één van de ouders. Samenvattend geven deze resultaten aan dat het ontstaan van *de novo* mutaties niet beperkt is tot de gametogenese maar dat een belangrijk deel van *de novo* mutaties post-zygotisch ontstaat. Genetisch mozaïcisme dat ontstaat door mutaties in gedurende de embryogenese komt veel voor en kan afhankelijk van het moment van ontstaan doorgegeven worden aan het nageslacht. Dit suggereert dat onze genomen veel dynamischer zijn dan eerder gedacht.

Het ontstaan van nieuwe mutaties in mensen is niet beperkt tot de vroege stadia van het leven and mutaties ontstaan ook na de geboorte en gedurende het volwassen leven. In **hoofdstuk 5** onderzoeken we de timing van nieuwe mutaties in weefsels gedurende het leven door middel van het onderzoeken van somatische mutaties in menselijke hematopoiese. In dit onderzoek richten we onze aandacht op mutaties die leiden tot positieve selectie van hematopoiese stamcellen die de uitbreiding van klonen met bepaalde mutaties voortdrijven. Deze mutaties staan bekend als klonale hematopoiese sturende mutaties en zijn vaak aanwezig als laaggradige mozaïeke mutaties in bloed. Om deze somatische mutaties te kunnen detecteren ontwikkelden we een test die twee zaken combineerde: toepasbaarheid op grote hoeveelheden mutaties en zeer grote gevoeligheid. Hiermee testen we het bloed van individuen met leeftijden tussen de 20 en 69 jaar op genetische veranderingen. Klonale hematopoiese sturende mutaties werden gevonden in het bloed van 3% van de individuen met een leeftijd tussen 20 en 30 jaar. Bovendien vonden we een exponentiële toename van de frequentie van klonale hematopoiese sturende mutaties in het bloed van meer dan 20% van de individuen met een leeftijd tussen de 60 en 69 jaar. De bevindingen van dit onderzoek ondersteunen het idee dat het ontstaan van somatische mutaties in stamcellen, zoals hematopoiese stamcellen, samen met positieve selectie en de klonale evolutie van stamcellen met mutaties, een universeel mechanisme is dat plaatsvindt gedurende het menselijk leven.

Het moment waarop nieuwe mutaties ontstaan is van invloed op de gevolgen die deze mutaties hebben. Dezelfde mutaties kunnen bijvoorbeeld

leiden tot volledig verschillende fenotypes, afhankelijk van het moment waarop de mutatie ontstaat, zoals aangetoond in **hoofdstuk 6**. Dit hoofdstuk verkent de functionele gevolgen van overeenkomende kiembaan en somatische mutaties in het gen *SETBP1*, resulterend in Schinzel-Giedion syndroom, een zeldzame ontwikkelingsaandoening en myeloïde leukemie, respectievelijk. In dit onderzoek vonden we overeenkomsten tussen de uiteindelijke effecten van identieke kiembaan en somatische mutaties in *SETBP1*, inclusief biochemische en cellulaire consequenties. Bovendien bleek het mutatie spectrum verschillend tussen beide aandoeningen, ondanks dat er overeenkomstige mutaties zijn tussen kiembaan en somatische *SETBP1* mutaties in Schinzel-Giedion syndroom en myeloïde leukemie. Dit komt voort uit het feit dat ondanks dat *SETBP1* mutaties met geringe functionele effecten veel worden waargenomen als kiembaan gebeurtenissen in Schinzel-Giedion syndroom, ze maar zelden worden gevonden als somatische gebeurtenissen in myeloïde leukemie. Dit suggereert dat er verschillende functionele grenswaarden bestaan voor kiembaan en somatische mutaties waarbij kwaadaardige tumoren uitsluitend worden aangedreven door zeer verstorende mutaties in *SETBP1*, terwijl prenatale ontwikkeling wordt verstoord in Schinzel-Giedion syndroom door sterke dan wel milde mutaties in dit gen.

Tot slot, biedt **hoofdstuk 7** een gedetailleerde discussie over de bevindingen van dit proefschrift, inclusief het gebruik van NGS voor het detecteren en bestuderen van *de novo* kiembaan en somatische mutaties en de bijdrage van timing van nieuwe mutaties voor hun consequenties.

Het werk in dit proefschrift laat zien dat mutaties continue ontstaan, tussen generaties maar ook gedurende het leven. Dit continue ontstaan van mutaties leidt tot buitengewone genetische diversiteit tussen individuen maar staat ook aan de basis van het ontstaan van genetisch verschillende cel populaties in een enkele persoon. Alhoewel het ontstaan van nieuwe mutaties een belangrijk biologisch proces is in mensen met een belangrijke rol binnen prenatale ontwikkeling, fysiologie en evolutie, dragen nieuwe mutaties ook bij aan verschillende vormen van humane ziektes, van zeldzame ernstige ontwikkelingsaandoeningen tot aandoeningen die op volwassen leeftijd optreden zoals kanker. In de afgelopen jaren heeft NGS een centrale rol gespeeld in het identificeren en bestuderen van nieuwe mutaties, en heeft daarmee de humane genetica en de vertaling van resultaten naar klinische toepasbaarheid voortgestuwd. Het werk in dit proefschrift ondersteunt dat naast het identificeren van mutaties, NGS ook kan worden gebruikt als een instrument om het moment van het ontstaan van mutaties te vinden om zo de dimensie tijd toe te voegen aan de interpretatie van mutaties en hun mogelijke gevolgen. *De novo* mutaties hebben bijvoorbeeld een verschillend risico op herhaling, afhankelijk van het precieze moment waarop ze zijn ontstaan, wat kan worden vastgesteld door nauwkeurige analyse met NGS technieken. Bovendien, heeft het moment van

ontstaan van een mutatie ook gevolgen voor het uiteindelijke fenotype omdat dezelfde mutatie betrokken kan zijn bij verschillende aandoeningen afhankelijk van het ontstaansmoment. Toekomstig werk gericht op de rol en timing van somatische en kiembaan mutaties zal ons beter inzicht geven in het effect van de uiting van een mutatie in de dynamische context van een cel, een weefsel en een volledig organisme.

Resumen en Español

Toda la información necesaria para el desarrollo, crecimiento, vida y reproducción de una persona se encuentra codificada en el genoma humano y se transmite de padres a hijos. Las diferencias en la secuencia de ADN entre individuos se conocen como variantes genéticas. La mayoría de las variantes genéticas en el genoma de cualquier persona proviene del genoma de sus padres, quienes heredan estas variantes a su descendencia. Sin embargo, una pequeña parte de las variantes genéticas de cualquier persona aparecen por primera vez en el genoma de este individuo y se consideran por lo tanto mutaciones nuevas, llamadas *de novo*. Al igual que cualquier mutación genética, las mutaciones *de novo* pueden ser benignas, neutrales o dañinas y juegan un papel importante en la evolución, diversidad y enfermedad en el ser humano. En esta tesis, he estudiado en qué momento a lo largo de la vida humana aparecen nuevas mutaciones y qué consecuencias tiene una mutación dependiendo del momento en el que aparece.

El **Capítulo 2** de esta tesis ofrece una revisión detallada de los mecanismos tras la aparición de nuevas mutaciones y su distribución en el genoma. Una mutación surge cuando se produce un daño en el ADN que no se repara de forma exitosa, lo cual puede suceder en cualquier momento y en cualquier célula del organismo. Como resultado, las mutaciones pueden aparecer a lo largo de toda la vida humana, desde mutaciones en el espermatozoide o el óvulo durante la formación de los gametos dando lugar a mutaciones *de novo* en la línea germinal de la progenie, hasta mutaciones en el tejido somático durante la vida adulta.

Las mutaciones nuevas son una causa frecuente de afecciones genéticas en el ser humano, en especial de enfermedades genéticas de aparición esporádica que causan alteraciones graves que afectan la adaptabilidad de un organismo o su fertilidad, como por ejemplo ciertas enfermedades pediátricas. El **Capítulo 3** de esta tesis es un estudio en el que identificamos mutaciones *de novo* en la línea germinal en el gen *THRA* como la causa de un síndrome de resistencia a la hormona tiroidea caracterizado por alteraciones en el desarrollo y crecimiento. Los resultados de este estudio sirven como ejemplo del papel que desempeñan las mutaciones *de novo* en las enfermedades pediátricas severas y de aparición esporádica. Identificamos las mutaciones responsables de este síndrome de resistencia a la hormona tiroidea por medio de la secuenciación de exomas, seguido de un análisis comparativo de variantes genéticas entre dos individuos sin parentesco, afectados por esta enfermedad. Este estudio subraya la importante contribución del desarrollo de Next Generation Sequencing (NGS) a nuestra comprensión del papel de las mutaciones *de novo* en las enfermedades del ser humano.

Aunque la mayoría de las veces las mutaciones *de novo* aparecen durante la formación de los gametos o antes de la primera división celular del cigoto, también pueden ocurrir después de esta primera división celular, dando lugar a mosaicismo genético. De hecho, algunas enfermedades del desarrollo humano son causadas exclusivamente por mutaciones *de novo* postcigóticas. En el **Capítulo 4**, usamos diferentes técnicas de NGS para estudiar un grupo de más de cien mutaciones *de novo* y determinar qué proporción de las cuales corresponde a mutaciones de origen postcigótico. Descubrimos que el 7% de las mutaciones *de novo* que aparentaban ser mutaciones de línea germinal, corresponden en realidad a mutaciones de origen postcigótico. Además, encontramos varias instancias en las que mutaciones identificadas en sujetos de nuestra cohorte como mutaciones *de novo* de línea germinal, fueron posteriormente encontradas en ADN extraído de la sangre de uno de los padres. Este hallazgo sugiere que, contrario a lo que pensábamos, estas mutaciones no ocurrieron *de novo* en el sujeto estudiado sino antes, durante el desarrollo embrionario del padre en el que la mutación fue detectada. En conjunto, nuestros resultados indican que la aparición de mutaciones *de novo* no se limita a la etapa de formación de gametos, sino que una porción considerable de las mismas ocurre de forma postcigótica en el embrión en desarrollo. Por lo tanto, el mosaicismo es un fenómeno común y, en función del momento en el que aparecen, las mutaciones postcigóticas pueden transmitirse a la siguiente generación. Estos hallazgos sugieren que nuestros genomas serían más dinámicos de lo que se habría considerado anteriormente.

La aparición de nuevas mutaciones no se limita únicamente a los estadios más tempranos de vida en el ser humano, sino que pueden seguir apareciendo a lo largo de toda la vida. En el **Capítulo 5**, usamos la hematopoyesis como modelo de estudio para investigar cuándo ocurren nuevas mutaciones en tejidos somáticos a lo largo de la vida humana. En este capítulo, nos enfocamos a estudiar mutaciones que promueven la selección positiva de células madre hematopoyéticas, causando así la expansión de clones de células madre hematopoyéticas con mutaciones. A las mutaciones que provocan este efecto se les conoce como mutaciones causantes de hematopoyesis clonal y su presencia se detecta en sangre frecuentemente como bajos niveles de mosaicismo. Para poder detectar la presencia de estas mutaciones en sangre, desarrollamos un método que combina alto rendimiento con secuenciación ultra-sensible y estudiamos individuos entre los 20 y los 69 años de edad. Detectamos mutaciones de este tipo en sangre en 3% de la población entre los 20 y 30 años de edad. Además, observamos que con la edad se produce un aumento exponencial en la frecuencia de estas mutaciones en sangre, llegando a sobrepasar el 20% de la población entre 60 y 69 años de edad. Los hallazgos de este estudio sugieren que la aparición de mutaciones somáticas en células madre y la subsiguiente selección positiva y evolución de clones de células con mutaciones es un mecanismo universal que ocurre a lo largo de toda la vida en el ser humano.

El momento de aparición de una mutación determina las consecuencias que puede llegar a tener esta mutación. Por ejemplo, mutaciones iguales pueden dar lugar a fenotipos completamente distintos, en función de cuándo ocurren, como mostramos en el **Capítulo 6**. Este capítulo explora las consecuencias funcionales de mutaciones en un gen llamado *SETBP1* dependiendo de si ocurren como mutaciones en la línea germinal causando el síndrome de Schinzel-Giedion, una enfermedad rara del desarrollo, o si ocurren como mutaciones somáticas como sucede en la leucemia mieloide. En este estudio identificamos similitudes a nivel bioquímico y celular, entre las consecuencias de mutaciones de línea germinal y mutaciones somáticas que se sobreponen. Además, a pesar del traslape, el espectro de mutaciones es distinto en cada condición. Esto se debe a que las mutaciones en *SETBP1* que tienen consecuencias funcionales leves, se observan con frecuencia en el síndrome de Schinzel-Giedion pero no tan frecuentemente en la leucemia mieloide. Las mutaciones que ocurren en la leucemia mieloide son casi exclusivamente mutaciones con consecuencias funcionales graves, mientras que en el síndrome de Schinzel-Giedion se detectan mutaciones con consecuencias funcionales tanto leves como intensas. Esto da pie a la posibilidad de que los umbrales de patogenicidad de las mutaciones somáticas involucradas en cáncer y de las mutaciones de línea germinal involucradas en enfermedades del desarrollo sean diferentes

Finalmente, el **Capítulo 7** contiene una discusión detallada de los hallazgos de esta tesis, incluyendo el uso de NGS para la detección de mutaciones nuevas tanto en la línea germinal como somáticas, así como para estudiar cómo el momento de aparición de una mutación influye sobre sus posibles consecuencias.

El trabajo presentado en esta tesis muestra que las mutaciones aparecen de forma constante entre una generación y la siguiente pero también a lo largo de toda la vida. Esta aparición constante de mutaciones está detrás de la extraordinaria diversidad genética entre individuos, pero es también el origen de la existencia de poblaciones de células genéticamente distintas en una misma persona. La aparición de mutaciones nuevas representa un fenómeno biológico crucial en el ser humano, con un rol en el desarrollo embrionario, en la fisiología y en la evolución. Sin embargo, también contribuye a la aparición de distintas formas de enfermedad, incluyendo graves enfermedades congénitas y enfermedades de inicio tardío como el cáncer. En los últimos años, el NGS ha tenido un papel crucial en la identificación y el estudio de nuevas mutaciones, promoviendo el avance de la investigación en el campo de la genética humana y permitiendo de forma temprana la traducción de resultados al mundo clínico. El trabajo contenido en esta tesis propone que, además de utilizarse para la detección de mutaciones, es posible emplear el NGS como una herramienta para identificar el momento de aparición de una mutación y así poder incluir la dimensión del tiempo en la interpretación de las mutaciones y de sus posibles consecuencias. Por ejemplo, el riesgo de recurrencia de las mutaciones *de novo*

difiere en función del momento exacto en el que ocurren, lo cual puede detectarse por medio de un análisis meticuloso basado en NGS. Adicionalmente, el momento de aparición de una mutación también puede moldear sus consecuencias, ya que la misma mutación puede estar involucrada en distintas enfermedades en función de su momento de aparición. A futuro, la investigación sobre el papel que desempeña el momento de aparición de las mutaciones en la línea germinal y de las somáticas, nos proporcionará una mejor comprensión del efecto que tienen en el contexto dinámico de una célula, un tejido y un organismo completo.

Acknowledgements

I owe this work to many people who helped and supported me throughout the years of my PhD, but also to many people who unknowingly set the foundations which allowed me to complete this thesis.

Alex, you have been an extraordinary mentor and friend during these years. I have thoroughly enjoyed the time we have spent working together and I have learned so much from you. I include solid scientific knowledge among the lessons learned, of course, but I particularly value the freedom you gave me to explore different projects and acquire new skills, both in Nijmegen and abroad. From our very first conversation on internship projects up until our last brainstorm session for future projects for after the PhD, your enthusiasm and passion for science has always been inspiring and contagious, constantly sparking intriguing scientific questions. From the bottom of my heart, thank you for all the time and energy you have invested, not only in supervising me in my PhD project, but also in guiding me with so much care in my first steps as a scientist.

Christian, your memories of life as a PhD student were still fresh in 2012, so I quickly turned to you as a sort of guru who would explain life to me with nonchalance. I'm thrilled to see the evolution of our academic lives over these years and fully appreciate the mark you've left in mine. I find your hard work and passion for science inspiring but, above all, I've always enjoyed the mix of healthy criticism and dry humor you usually sprinkle on our work. I deeply value the efforts you have invested in teaching and guiding me, as well as the time we have spent together (including your patient but infructuous initial attempts to teach me programming, discussions of all sorts and youthful tomfoolery at conferences worldwide). Thank you for your support and help as a supervisor and as a friend over these last years.

Joris, I have always felt very lucky to work with you; you have pioneering vision, a talent to deal with people and, still, you remain approachable, with a great sense of humor and never hesitate to give a Dutch compliment when needed. I truly consider it's been an honor to work in a group with such high-quality science and cutting-edge technology. Although your pragmatism often met stubborn resistance from my side, retrospectively I realize that you were right about many things concerning my project (possibly, about almost everything). I'm happy to think that I've picked up some of that along the way. I value very much the freedom you gave me to lead my project and explore different options, while jumping in with new ideas and proposals when you felt it was needed. I will certainly miss your good humor, relaxed approach and Spanish greetings.

Han, I want to thank you for your guidance and for all the time invested in supervising me. I always found our discussions and brainstorm sessions fascinating and came out of each one of them with great ideas or interesting papers to read. Thank you.

It has been delightful to spend the last few years working at our department, in an environment bubbling with science and yet welcoming and *gezellig*. **Marloes**, you are a very important contributor to this thesis and worked very hard on different projects included. For this, I'm very thankful and glad to have you by my side at my defense. I appreciate your fun sense of humor, curiosity, frankness and drive and look forward to sharing with you the end of this cycle. **Michael**, I'm of course grateful for your work on the postzygotic mutations project, but also for the reliability of your good humor, your friendliness and your enthusiastic participation in my bicycle thievery and other chaotic adventures. **Petra**, you are full of energy of the good kind and wherever I go, I know I will miss you and your positivity! **Lisenka**, scientist extraordinaire, I am continuously astonished with the precision of your vast scientific knowledge and capacity to calmly multitask. **Peer**, we shared a lot of PhD experiences, sitting next to each other for years, migrating from office to office. Starting the day with a coffee and a chat with you and your optimism is a ritual I will definitely miss. Although you joined our group later, **Manon**, **Sinje** and **Gaël**, I appreciate the good energy you brought to our team that made it such a pleasant environment to work in. **Irene** and **Konny**, it was lovely to work and share nice moments with you.

Without the magic of bioinformaticians, my PhD project would simply not have developed. **Maartje**, I am in awe that there was never a single request that you couldn't complete, always making it seem like it wasn't a big deal. Thank you for all your hard work over these last years. **Stefan**, I appreciate your concern for my wellbeing, making sure I'd join for much-needed lunch during the last phase of my PhD. I'm grateful for this and all the cup-tumbling cockatoos we've shared. **Jakob**, **Laurens**, **Jayne** and **Djie**, I prize very much the valuable input you gave during scientific discussions as well as your good humor and friendly approach. **Nienke**, **Marisol**, **Rick**, **Steven** and **Dimitra**, I appreciate your hard work to keep all things bioinformatics up and running.

I must say that without the magic of clinicians, my PhD project would not have gone very far either. **Bregje**, although I think it took a long and winded road to completion, I'm proud of our paper and it would not have been possible without your dedication and careful coordination with clinicians across the world. I really appreciate your supervision and work on this project. **Carlo**, **Bert** and **Sandra**, thank you for interesting scientific discussions and collaborations on different projects.

I am also grateful to the **patients with *SETBP1* mutations** and **their parents**, who assisted us in many ways with our research. I often found your efforts and interest in our research a strong motivation to go on. Thank you.

Pela, I can't think of a better companion to have worked on the Sisyphean task that was our *SETBP1* project. Your sense of humor and perseverance helped keep morale high in face of adversity and botched Western blots. You taught me a lot about grown up life in academia and I really value the time we spent working together. I am also very grateful to all the members of the Language and Genetics group, especially **Sara**, **Sarah** and **Elliot** and, of course, **Simon Fisher** for his enthusiasm in this project and patience.

I am grateful for the opportunity that Jay Shendure gave me to visit his lab and learn from the mind-blowing science that seems to be carried out almost effortlessly by his research group. This was truly inspiring and influenced my research, particularly in the last year of my PhD. **Jay**, you are an amazing scientist and a caring mentor, I am very thankful for this chance. **Greg**, I am truly grateful for your immense efforts, patience and good humor during the time we worked together in Seattle. I would like to thank all the members of the Shendure lab, who took the time to guide and teach me and especially **Molly**, **Andrew**, **Ron** and **Charlie** who made me feel at home during my stay in Seattle.

Nehir and **Hilal**, thank you for your hard work and great results, some of which are included in this thesis. I'm proud to see how you have gone on with your academic careers and hope you learned as much from your time with us as I did from you.

I am thankful to the members of the cancer group past and present but especially **Roland**, **Ad**, **Marjolijn Ligtenberg**, **Nicoline**, **Marjolijn Jongmans**, **Richarda**, **Robbert**, **Ingrid**, **Marc**, **Jiangyan**, **Junxiao**, **Simon**, **Eveline** and **Lilian**. You welcomed me to join your meetings at the beginning of my PhD, which helped me orient my project.

Lisette, thank you for patiently teaching me immunofluorescence and microscopy, my first steps in functional work in Nijmegen. **Erwin**, I am grateful for the insightful discussions at the beginning of my PhD. I am also thankful towards **Saskia** and the staff from the cell culture facility for their help with cell lines and DNA samples throughout my PhD.

Gerardo Gamba, siempre voy a sentirme agradecida de las oportunidades que me has dado y que me han abierto el mundo. Fuiste mi primer mentor en el mundo de la ciencia… ¡mira la huella que me dejó el artículo de secuenciación de exomas que nos compartiste a mediados de 2010! También quiero agradecerle a **Norma Vázquez** que con infinita paciencia me enseñó los

fundamentos de biología molecular. **Jasper Mullenders**, I appreciate the patience with which you taught me a variety of laboratory techniques but I am especially grateful towards you for teaching me that there are no such things as black boxes in science, an invaluable lesson for my PhD and future adventures in academia.

I shared beautiful experiences with the organizers and participants of the Radboud DaVinci Challenge, **Sami**, **Tom**, **Michiel**, **Arjan**, **Dennis**, **Jeroen**, **Peter** and **Tong**. Huge thanks to **Ellen** and to **Anja** for putting the program together and for inspiring discussions that have resonated more than one might imagine.

I would also like to thank all collaborators with whom we worked together during the last years, as well as all members of the Human Genetics department.

Many people contributed to this thesis, not by the means of brainstorms and PCRs, but by keeping life in balance and providing well-timed and sometimes much-needed support. From the bottom of my little heart, thank you all.

First of all, my gang from the hospital and all affiliated friends, with whom we've shared broodjes, beers, beards, brunch, bachelor and bachelorette parties, births of babies and much more. **Anchel**, **André**, **Angela**, **Antoine**, **Bart**, **Benny**, **Cindy**, **Davide**, **Ganesh** (and **Ramya**), **Katia**, **Marco**, **Markus**, **Nicco**, **Nuria**, **Pedro**, **Philipp**, **Sandhya**, **Simone**, **Sip**, **Stefania** and **Till**… I am glad we walked a part of the path together and hope we'll go on for much longer.

Blending Voices gave me a space to play, share and create which I appreciated at all times. I am grateful to all Blenders but especially **Alexandra**, **Ana**, **Bart**, **Daria**, **Dries**, **Elena**, **Giuliana**, **Johannet**, **Marc** and **Ruud**. Thank you for the beautiful moments we shared and everything each of you taught me.

To my Dutch friends **Wytze**, **Jeftha**, **Jamie**, **Simon v.R.** and **Dieter**, thank you for adopting me and providing experienced guidance in my mildly successful attempt at life as a Dutchwoman. I cherish the unique moments I shared with each of you. Your friendship made me feel rooted in Nijmegen and I consider it an invaluable experience that I take along with me. Hartelijk bedankt!

Alejandro, gracias por hacer de nuestra compartida existencia como huérfanos en el noroeste del Pacífico una experiencia positiva, por no mencionar las infusiones de humor y música durante la escritura de tesis. PAIS, que te toca!

Vicky, hemos compartido mucho: cenas, fiestas de cumpleaños y largas conversaciones aderezadas con café, pero también apoyo mutuo cuando la vida se tuerce. Gracias por tu bonita amistad durante todos estos años, has sido una

gran maestra y me alegro mucho de tenerte a mi lado en este momento importante.

Anchel, gracias por tu apoyo, compañía y todos los buenos momentos que compartimos en este trecho del camino. Siempre me prestaste una oreja (y a veces incluso un hombro) y me animaste a seguir adelante, dándome buenos consejos basados en tu experiencia. Te lo agradezco con muchísimo cariño.

Daria, mi amiga cronopia, gracias por tu hermosa amistad y por haberme dado a conocer, de forma inesperada, el amor por el teatro. Siempre guardaré en mi corazón los recuerdos de los momentos maravillosos que compartimos. Obviamente, corriendo libres por la casa, como debe ser.

Rodrigo, hemos vivido viajes y vicisitudes desde que nos conocemos: nuestra existencia chilanga, la aventura de venir a Holanda y aprender a andar en bicicleta y usar la calefacción. Aunque ahora hemos creado una vida por separado, eres lo más parecido a familia que tengo en Nijmegen. Gracias por la última década y media y por venir a levantarme cuando ha hecho falta.

Julia, mi amiga, you always appear at the right moment, bring your wonderful and transformative energy and leave when your work is done. Gracias por tu amistad todos estos años y las aventuras que hemos vivido juntas.

David, you appeared in the middle of my frantic thesis writing and decided to stay, calmly watering me with coffee and treating me to cat swarms until the very end. Pure bliss. Thank you for being a serene counterweight to my chaos and showing me that even though I could be alone, I didn't have to and that I could feel at home with you.

A Gustavo Cerati, sus colores santos y amor amarillo que fueron mis compañeros durante largas noches en los últimos meses y me inspiraron para ir a conocer la Ciudad de la Furia por mí misma.

Finalmente, le estoy agradecida a mi extraordinaria familia, que siempre ha tolerado mis ausencias de casa sin dejar de darme viento para volar. **Bea**, nunca me dejas de asombrar con tus maravillosas inverosimilitudes y nuestra complicidad que sobrevive a la distancia. En más de una ocasión recibí un mensaje tuyo, repleto de esa aleatoriedad que te caracteriza, en un momento en el que lo necesitaba. Gracias ma petite bibichoune. **Mamá** y **Papá**, gracias por absolutamente todos los esfuerzos que han hecho por mí y que han creado el camino para que pudiera llegar hasta aquí. Desde los libros que mi padre me traía al volver del trabajo cuando era chica hasta las últimas visitas de mi madre a Nijmegen tan solo meses antes de terminar la tesis, ha sido un largo recorrido que no podría haber emprendido y completado de no ser por ustedes. Mil gracias.

List of publications

Acuna-Hidalgo R., Sengul H., Steehouwer M., van de Vorst M., Vermeulen S.H., Kiemeney L.A.L.M., Veltman J.A., Gilissen C. & Hoischen A. Somatic driver mutations in age-related clonal hematopoiesis by ultra-sensitive sequencing. *Submitted.*

Acuna-Hidalgo R.*, Deriziotis P.*, Steehouwer M., Gilissen C., Graham S.A., van Dam S., Hoover-Fong J., Telegrafi A.B., Destree A., Smigiel R., Lambie L.A., Kayserili H., Altunoglu U., Lapi E., Uzielli M.L., Aracena M., Nur B.G., Mihci E., Moreira L.M., Borges Ferreira V., Horovitz D.D., da Rocha K.M., Jezela-Stanek A., Brooks A., Reutter H. Cohen J.S., Fatemi A., Smitka M., Grebe T., Di Donato N., Deshpande C., Vandersteen A., Marques Lourenço C., Dufke A., Rossier E., Andre G., Baumer A., Spencer C., McGaughran J., Franke L., Veltman J.A., De Vries B.B.A., Schinzel A., Fisher S.E., Hoischen A. and van Bon B.W. Overlapping SETBP1 gain-of-function mutations in Schinzel-Giedion syndrome and hematologic malignancies. PLoS Genetics 2017 Mar 27;13(3):e1006683.

Acuna-Hidalgo R., Veltman J.A., Hoischen A. New insights into the generation and role of de novo mutations in health and disease. Genome Biol. 2016 Nov 28;17(1):241.

Acuna-Hidalgo R., Bo T., Kwint M.P., van de Vorst M., Pinelli M., Veltman J.A., Hoischen A., Vissers L.E., Gilissen C. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. Am J Hum Genet. 2015 Jul 2;97(1):67-74.

Tylki-Szymańska A.*, **Acuna-Hidalgo R.*,** Krajewska-Walasek M., Lecka-Ambroziak A., Steehouwer M., Gilissen C., Brunner H.G., Jurecka A., Różdżyńska-Świątkowska A., Hoischen A., Chrzanowska K.H. Thyroid hormone resistance syndrome due to mutations in the thyroid hormone receptor α gene (THRA). J Med Genet. 2015 May;52(5):312-6.

* Shared first authorship

Acuna-Hidalgo R.*, Schanze D.*, Kariminejad A.*, Nordgren A.*, Kariminejad M.H., Conner P., Grigelioniene G., Nilsson D., Nordenskjöld M., Wedell A., Freyer C., Wredenberg A., Wieczorek D., Gillessen-Kaesbach G., Kayserili H., Elcioglu N., Ghaderi-Sohi S., Goodarzi P., Setayesh H., van de Vorst M., Steehouwer M., Pfundt R., Krabichler B., Curry C., MacKenzie M.G., Boycott K.M., Gilissen C., Janecke A.R., Hoischen A., Zenker M. Neu-Laxova syndrome is a heterogeneous metabolic disorder caused by defects in enzymes of the L-serine biosynthesis pathway. Am J Hum Genet. 2014 Sep 4;95(3):285-93.

Acuña R., Martínez-de-la-Maza L., Ponce-Coria J., Vázquez N., Ortal-Vite P., Pacheco-Alvarez D., Bobadilla N.A., Gamba G. Rare mutations in SLC12A1 and SLC12A3 protect against hypertension by reducing the activity of renal salt cotransporters. J Hypertens. 2011 Mar;29(3):475-83.

Ramos-Rivas M., Rojas-Velasco G**., Acuña-Hidalgo R.**, Márquez-Valero O.A., Arellano-Bernal R.H., Castro-Martínez E. [Paraneoplastic limbic encephalitis: a difficult-to-diagnose condition]. Rev Neurol. 2009 Mar 16-31;48(6):311-6.

* Shared first authorship

Curriculum vitae

Rocío Acuña Hidalgo was born in Buenos Aires, Argentina on February 14th 1985. Her childhood was spent wandering around the world until her family decided to settle in Mexico City, Mexico. After graduating from the Lycée Franco-Mexicain in 2003, Rocío began her studies in medicine at La Salle University in Mexico City. As part of her medical degree, she completed one year of mandatory social service performing basic research under the supervision of Dr. Gerardo Gamba at the *Instituto Nacional de Ciencias Médicas y Nutrición*, where she characterized the functional consequences of rare mutations in sodium co-transporters linked to low blood pressure. Rocío received her Medical Degree with Honors in 2010 and, shortly after, was awarded a Huygen's scholarship from the Dutch Minister of Education, Culture and Science to study a Master's in Molecular Mechanisms of Disease at Radboud University in Nijmegen, The Netherlands. Interested in Next Generation Sequencing, she completed an internship at the Department of Human Genetics under the supervision of Joris Veltman and Alexander Hoischen where she performed whole exome sequencing to identify disease-causing genes in severe developmental disorders and prenatal lethal phenotypes. She was subsequently awarded a personal grant from the Radboud University Medical Centre (UMC) to pursue a PhD and, after receiving her Master's degree *cum laude*, Rocío joined Joris Veltman's group in 2012. Here, she worked on several projects studying the timing of novel mutations and the genetic overlap between developmental disorders and cancer. Rocío took part in the first edition of the Radboud DaVinci Challenge, a talent development programme from Radboud UMC. In 2015, Rocío visited Jay Shendure's research group at the University of Washington in Seattle to learn Saturation Genome Editing, a novel technique for massive parallel mutagenesis. For this, she was awarded a Frye stipend, a travel grant for promising female scientists. Rocío received the Young Investigator Award for Outstanding Science at the European Society for Human Genetics conference in 2016, after which she completed her PhD the following year.

RIMLS portfolio

Activity	ECs	Year(s)
Courses & Workshops		
Graduate course (Radboud Institute for Molecular Life Sciences, RIMLS)	2	2013
Academic writing (Radboud Unversity)	3	2013-14
Technical forums and workshops (RIMLS)	0.9	2013-16
Seminars & lectures		
Seminars and lectures (RIMLS)	3.7	2012-17
Seminars and lectures (Washington University, Seattle, WA, USA)	0.3	2015
National congresses		
New Frontiers in Genetics – Nijmegen, NL	0.5	2012
Rolduc Genetica retraite* – Kerkrade, NL	0.75	2013
New Frontiers in synthetic life – Nijmegen, NL	0.5	2013
Rolduc Genetica retraite** – Kerkrade, NL	0.75	2014
RIMLS PhD retreat** – Wageningen, NL	0.75	2014
New Frontiers in regenerative medicine – Nijmegen, NL	0.5	2014
RIMLS PhD retreat** – Veldhoven, NL	0.75	2015
RIMLS PhD retreat* – Veldhoven, NL	0.75	2016
International congresses		
European Society for Human Genetics – Paris, France	1	2014
European Society for Human Genetics** – Milan, Italy	1.5	2015
Genomics of Rare Disorders* - Hinxton, UK	1.25	2015
Rare Diseases, crossing borders together§ - Białobrzegi, Poland	1.5	2015
American Society for Human Genetics* – Baltimore, USA	1.75	2015
European Society for Human Genetics* – Barcelona, Spain	1.75	2016
American Society for Human Genetics** - Vancouver, Canada	1.5	2016
European Society for Human Genetics* – Copenhaguen, Denmark	1.75	2017
Student supervision		
Nehir Kurtas –Master's in Molecular mechanisms of disease	2	2014
Hilal Sengül– Master's in Molecular mechanisms of disease	2	2016
Other activities		
Literature discussion (RIMLS)	3.5	2012-16
Literature discussion (Washington University, Seattle, WA, USA)	0.5	2015
Member of the RIMLS PhD committee	1	2013-15
Chair of Genomic Disorders meeting	2	2014-15
Total	**38.15**	

* oral presentation; ** poster presentation; § invited presentation

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.
T.S. Elliot